BÜRO FÜR TECHNIKFOLGEN-ABSCHÄTZUNG
BEIM DEUTSCHEN BUNDESTAG

Katrin Gerlinger

Data-Mining – gesellschaftspolitische und rechtliche Herausforderungen

Endbericht zum TA-Projekt

Juni 2022
Arbeitsbericht Nr. 203

Katrin Gerlinger

Data-Mining – gesellschaftspolitische und rechtliche Herausforderungen

Endbericht zum TA-Projekt

TAB-Arbeitsbericht Nr. 203

Büro für Technikfolgen-Abschätzung
beim Deutschen Bundestag
Neue Schönhauser Straße 10
10178 Berlin

Telefon: +49 30 28491-0
E-Mail: buero@tab-beim-bundestag.de
Web: www.tab-beim-bundestag.de

2022

Umschlagbild: nicoelnino/123RF

ISSN-Internet: 2364-2602

Das Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag (TAB) berät das Parlament und seine Ausschüsse in Fragen des wissenschaftlich-technischen Wandels. Das TAB wird seit 1990 vom Institut für Technikfolgenabschätzung und Systemanalyse (ITAS) des Karlsruher Instituts für Technologie (KIT) betrieben. Hierbei kooperiert es seit September 2013 mit dem IZT – Institut für Zukunftsstudien und Technologiebewertung gGmbH sowie der VDI/VDE Innovation + Technik GmbH.

Inhalt

Zusammenfassung 7

1 Einleitung 27

2 Data-Mining aus analytisch-technischer Sicht 35

2.1 Data-Mining – was ist das? 35

2.2 Daten: Formen, Strukturen und Bereitstellung 42

2.2.1 Wesensmerkmale und Formen 42

2.2.2 Datenspeicherung und -bereitstellung: von Datenbanken bis Systemarchitekturen 46

2.3 Data-Mining als Prozess: Schritte, Verfahren, Ergebnisse 54

2.3.1 Spezifikation der Untersuchungsaufgabe und Datenaufbereitung 54

2.3.2 Datenanalytische Verfahren 56

2.3.3 Ergebnisprüfungen 61

2.3.4 Weiterverwendung von Data-Mining-Ergebnissen 69

3 Rechtliche und normative Aspekte 73

3.1 Datenbezogene Grundstrukturen 74

3.2 Umgang mit nichtpersonenbezogenen Daten: Beispiel Geodaten und nationale Geodateninfrastruktur 82

3.3 Umgang mit personenbezogenen Daten 96

3.3.1 Von Datenverarbeitung betroffene Personen und deren Rechte 96

3.3.2 Grundsätze und Pflichten bei der Datenverarbeitung 101

3.3.3 Grundrechteschützende Maßnahmen 108

3.3.4 Das Forschungsprivileg – ein Türöffner für Data-Mining 119

3.3.5 Daten mit Bezug zu Personengruppen – (k)eine Sonderkategorie 124

3.4 Umgang mit Data-Mining-Ergebnissen 126

3.4.1 Informationen 127

3.4.2 Algorithmen und Software 128

3.4.3 Rechtsunsicherheiten und Entwicklungsinitiativen 130

4 Data-Mining in der Medizin 137
4.1 Medizinische Daten: rechtliche und technische Aspekte 137
4.1.1 Ärztinnen und Ärzte: Aufgaben, Pflichten, Aktenführung 138
4.1.2 Medizinische Einrichtungen: Organisation und Datenverwaltung 142
4.1.3 Medizinische Primärdaten 144
4.1.4 Aufbereitete medizinische Datenbestände 153
4.1.5 Gesamteinschätzung Datenzugänglichkeit 158
4.2 Medizinprodukte zur Generierung und Analyse medizinischer Daten 160
4.2.1 Prüfung und Bewertung der Sicherheit, Leistung und Gesundheitseffekte 162
4.2.2 Integration in die medizinische Versorgung 163
4.2.3 Haftung und Schadensausgleich 165
4.3 Data-Mining-Anwendungsbeispiele 168
4.3.1 Risikoklassifikation und medizinisches Scoring 169
4.3.2 Bilderkennung bei der Mammografie 171
4.3.3 Interpretation genetischer Daten für die Therapieplanung 179
4.3.4 Medizinische Assistenzsysteme 187
4.4 Administrative Daten: Basis von Geschäftsprozessen 196
4.4.1 Daten zur Leistungsabrechnung 197
4.4.2 Daten zur Qualitätssicherung der medizinischen Versorgung 202
4.4.3 Daten für die amtliche Statistik 205
4.4.4 Gesamteinschätzung der analytischen Potenziale administrativer Daten 205
5 Data-Mining im Gesundheitssystem 207
5.1 Sozialdaten: Rechtsrahmen der Verarbeitung 207
5.2 Das Institut für das Entgeltsystem im Krankenhaus: Daten und Analytik 209
5.2.1 Die stationäre Leistungsvergütung als lernendes System 209
5.2.2 Daten und deren Weiterverwendungsmöglichkeiten 211
5.2.3 Die DRG-Systemfortschreibung – ein Data-Mining-Prozess? 212
5.2.4 Einschätzung 213

5.3 Kassenärztliche Vereinigungen: Daten und Analytik 216
5.3.1 Kennnummern, Register und Informationssysteme der Kassenärztlichen Vereinigungen 217
5.3.2 Daten zur Qualitätssicherung und Data-Mining-Potenziale 219
5.3.3 Leistungsdaten: Prüfung, Verwendung, Weiterleitung 221
5.3.4 Sekundärnutzung von Leistungsdaten: Data-Mining-Beispiel »Verbreitung multiresistenter Erreger« 223
5.3.5 Einschätzung 227
5.4 Arzneimittelversorgung: Akteure, Daten und deren Verwendungsmöglichkeiten 228
5.4.1 Vereinigungen, Register, Informationssysteme 229
5.4.2 Apothekenrechenzentren 230
5.4.3 Das Deutsche Arzneiprüfinstitut 232
5.4.4 Exkurs: kommerzielle Datenweiterverwendung – ein zulässiges Geschäftsmodell? 233
5.4.5 Einschätzung 235
5.5 Gesetzliche Krankenkassen: Daten und Analysemöglichkeiten 236
5.5.1 Aufgaben, Strukturen, Datenbestände 236
5.5.2 Daten aus der Leistungsabrechnung: Bestandteile, Haltung, Mehrfachnutzung 241
5.5.3 Sekundärnutzung von Leistungsdaten: Data-Mining-Beispiel »Pharmakovigilanz« 249
5.6 Fazit 265
6 Gesamtfazit und Handlungsoptionen 269
6.1 Fazit 269
6.2 Allgemeine Handlungsoptionen 269
6.3 Handlungsoptionen, die sich aus dem Vergleich der Fallbeispiele ableiten lassen 270
7 Literatur 277
7.1 In Auftrag gegebene Gutachten 277
7.2 Weitere Literatur 277

8 Anhang 295
8.1 Data-Mining im Urheber- und Leistungsschutzrecht 295
8.2 Abbildungen 303
8.3 Tabellen 303
8.4 Kästen 303
8.5 Abkürzungen 303

Zusammenfassung

Datenanalysetechniken, die in den kontinuierlich größer werdenden Datenbeständen (neue) Strukturen erkennen, werden einerseits vielfältige Innovationspotenziale zugeschrieben, weil wichtige Erkenntnisse gewonnen, Prozessabläufe verbessert sowie Geschäftsideen und Informationsdienste entwickelt werden können. Andererseits werden auch Bedenken geäußert: Die Spanne reicht von intransparenten Abläufen über ungleiche Verwertungsmöglichkeiten abgeleiteter Informationen bis zu Veränderungen des gesellschaftlichen Miteinanders und dem Verlust der Privatheit. Bei der Aufgabe, diese technologische Entwicklung für die Allgemeinheit sinnvoll zu gestalten, ergeben sich vielfältige Herausforderungen, u. a. bezüglich der notwendigen Datenbereitstellung, der Konkretisierung von Möglichkeiten und Grenzen der Datenverwendung, des Umgangs mit den Ergebnissen, der Ausgestaltung von Verantwortungs- und Haftungsfragen sowie damit verbundener Finanzierungs- und Geschäftsmodelle. Nicht alle Herausforderungen sind fundamental neu, denn Daten werden seit langem erfasst und analysiert. In Anbetracht der kontinuierlich größer werdenden Datenbestände, deren vielfältigen Verknüpfungsmöglichkeiten und der analytisch-technischen Entwicklungen erscheint eine Auseinandersetzung mit den Möglichkeiten und Grenzen komplexer Datenanalysen und den damit einhergehenden Folgen jedoch erforderlich.

Der Ausschuss für Bildung, Forschung und Technikfolgenabschätzung des Deutschen Bundestages hat das Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag (TAB) mit einer Untersuchung zum Thema Data-Mining beauftragt, die mit diesem Bericht abgeschlossen wird. Im Zentrum der Betrachtung stehen Datenbestände, die im Rahmen öffentlicher Aufgaben erhoben und verarbeitet werden, Analysetechniken, die mit dem Begriff Data-Mining assoziiert werden, sowie das rechtliche Fundament, das Möglichkeiten und Grenzen der Datenanalytik teils allgemein, teils bereichsbezogen definiert. Vertiefend betrachtet werden medizinische und gesundheitssystemische Anwendungsbereiche. Dort werden komplexen Datenanalysen regelmäßig besondere Anwendungspotenziale unterstellt, aber auch Defizite in der Digitalisierung zahlreicher Prozessabläufe und folglich bei der Datenbereitstellung attestiert.

Ziel dieses Berichts ist es, den Oberbegriff Data-Mining aus unterschiedlichen Perspektiven zu erschließen und in seiner Vielschichtigkeit darzustellen. Damit soll das Verständnis der Möglichkeiten und Grenzen komplexer Datenanalysen erhöht werden. Anhand von unterschiedlichen Anwendungsbeispielen werden derzeitige Möglichkeiten und Herausforderungen in medizinischen und gesundheitssystemischen Kontexten veranschaulicht.

Data-Mining aus analytisch-technischer Sicht

Seit den 1990er Jahren wird der Begriff Data-Mining von Datenanalyst/innen mit der Anwendung mathematisch-statistischer Verfahren assoziiert, die Strukturen und Muster in verfügbaren Datenbeständen erkennen und entsprechende Informationen liefern. Wenn man Fragen nach den mit den Analysetechniken verbundenen gesellschaftlichen Chancen und Herausforderungen nachgehen will, sollte man nicht nur den unmittelbaren Einsatz von strukturerkennenden statistischen Verfahren (*Data-Mining im engeren Sinn*), sondern den gesamten Prozess der Informationsgewinnung aus Datenbeständen betrachten (*Data-Mining im weiteren Sinn)*. Data-Mining kann als zweckgebundene Datenanalytik verstanden werden, wobei konkrete Fragestellungen die jeweilige Untersuchungsaufgabe spezifizieren: Welche Faktoren/Merkmale indizieren erhöhte Erkrankungsrisiken? Welche Datenobjekte sind sich ähnlich, können zusammengefasst, gruppiert oder bestehenden Klassen zugeordnet werden? Welche Abweichungen in den Daten deuten auf Besonderheiten hin (u. a. Fehler, Risikogebiete, Straftaten)? Je nach Aufgabe und eingesetzten Verfahren werden statistische Kennziffern ermittelt, Parameter von Modellen angepasst oder Entscheidungsregeln abgeleitet, die im jeweiligen Kontext verallgemeinerbar sein sollten und beispielsweise zu Verfahren zur Bilderkennung, zum Scoring von Objekten oder zur Prognose von Sachverhalten weiterentwickelt werden können. Folgende Schritte gehören zum Data-Mining-Prozess:

- *Definition der Aufgabe als mathematisches Problem* (Suche nach Ähnlichkeiten oder Unterschieden/Ausreißern in Datenbeständen, Klassifikationen/Gruppierung von Objekten, Ableitung von Regeln, Modellierung);
- *Datenauswahl und -aufbereitung* (Prüfung der Dateneignung, Fehlerbereinigungen, Umrechnungen, Erstellung von Analyse-/Trainingsdatensätzen);
- *Datenanalyse* (je nach Aufgabe und Datenform kommen unterschiedliche Verfahren in Betracht, einige gibt es seit Jahrzehnten [z. B. Cluster-/Regressions-/Assoziationsanalysen], andere wurden erst durch die erheblichen Steigerungen der Rechenleistung der letzten Jahre anwendungsreif [z. B. Bayes'sche oder künstliche neuronale Netze]);
- *Ergebnisvalidierung* (verfahrensintern mittels spezifischer Kennziffern sowie extern durch Vergleiche oder Beweise).

Data-Mining wird meist mit Analysen bereits vorhandener Datenbestände in Verbindung gebracht, die neu verknüpft oder zu neuen Zwecken weiterverwendet werden. Damit ist die originäre Datenerfassung dem Data-Mining-Prozess zwar vorgelagert, gleichwohl ist ein Verständnis des Erhebungskontextes wichtig, u. a. um die Eignung für die jeweilige Untersuchungsaufgabe einschätzen, Auffälligkeiten von Fehlern unterscheiden oder Ergebnisse interpretieren zu können. Es gibt unterschiedliche Meinungen, wie weit der Data-Mining-Prozess zu fassen ist, welche Schritte als vor- oder nachgelagert gelten.

Ein hoher Automatisierungsgrad ist bei der unmittelbaren Anwendung mathematisch-statistischer Verfahren auf einen aufbereiteten, standardisierten (Trainings-)Datensatz möglich (Data-Mining im engeren Sinn), weil Algorithmen alle analytischen Schritte definieren, die maschinell ausgeführt werden. Bei der Aufgabendefinition, der Datenaufbereitung und der Ergebnisprüfung sind nach wie vor vielfältige menschliche Tätigkeiten erforderlich. Meist werden in mehreren Schleifen (Zwischen-)Ergebnisse geprüft, das analytische Vorgehen angepasst sowie Daten hinzugezogen oder ausgeschlossen, bis die Resultate als ausreichend valide angesehen werden. Validitätsanforderungen variieren je nach Anwendungsbereich und möglichen Folgen.

Datenanalytische Vorgehensweisen sind seit jeher ein Kernelement wissenschaftlichen Arbeitens, jedoch längst nicht mehr auf diesen Bereich beschränkt. Datenbasierte Regeln und angepasste Modelle werden teilweise in gewerblichen Kontexten zu digitalen Werkzeugen und Informationsdiensten (algorithmische Entscheidungs[unterstützungs]systeme) weiterentwickelt oder der gesamte Prozess gewerblich realisiert.

Aus technischer Sicht benötigt man für Data-Mining Daten(bestände), analytische Verfahren sowie Hard- und Softwarearchitekturen, durch die Ressourcen bereitgestellt und Prozesse effizient realisiert werden können. Für die maschinelle Verarbeitung müssen Daten eine gewisse Struktur haben. Wenn standardisierte Terminologien und Codierungen sowie normierte Formate verwendet werden, können verschiedene Daten(sätze) verknüpft und erweitert werden. Metadaten geben Auskunft, welche Codierungen und Formate jeweils verwendet wurden. Diese Standardisierung gespeicherter Daten ist für Data-Mining-Aktivitäten hochrelevant und oft aufwendig.

Unter dem Oberbegriff Data-Mining werden vielfältige strukturerkennende mathematisch-statistische Verfahren subsummiert. Klassische Verfahren (z. B. für Regressions- oder Clusteranalysen) stellen Strukturen anhand von Parametern, Formeln oder Entscheidungsbäumen explizit dar (symbolische Verfahren). Künstliche neuronale Netze sind eine besondere Form mathematisch-statistischer Modelle, die an Trainingsdaten angepasst werden und maschinell lernen, definierte Aufgaben zu lösen, ohne Entscheidungsregeln darzustellen (subsymbolische Verfahren). Unter Analyst/innen gibt es unterschiedliche Meinungen, ob das Training künstlicher neuronaler Netze als Data-Mining aufgefasst werden sollte oder eine eigene Kategorie bildet. Die unterschiedlichen analytischen Ansätze können als sich ergänzende Möglichkeiten zur datenbasierten Lösung definierter Untersuchungsaufgaben aufgefasst werden. Jedes Verfahren hat eigene Stärken, Schwächen und Grenzen. In der Summe bilden sie ein umfangreiches methodisches Repertoire. Erhebliches Fachwissen ist erforderlich, um das für eine Untersuchungsaufgabe und einen verfügbaren Trainingsdatensatz passende Verfahren auszuwählen, deren Ergebnisse einzuordnen und zu bewerten.

Data-Mining lässt sich sowohl mit lokalen, weitgehend geschlossenen Hard- und Softwaresystemen (Data Warehouses) als auch mit dezentral vernetzten Systemen (Cloudstrukturen) realisieren, beide koppeln Datenspeicher mit Analysetools. Die Abläufe in Data Warehouses lassen sich tendenziell besser kontrollieren. Cloudstrukturen haben eine größere Ressourcenflexibilität für die Datenspeicherung und -verarbeitung. Beide Ansätze ergänzen sich oftmals.

Rechtliche Grundstrukturen

Der Begriff Data-Mining findet derzeit über das Urheberrecht Eingang in das Rechtssystem. Er wird dort definiert als automatisierte Analyse von digital vorliegenden Daten, mit deren Hilfe Informationen, u. a. über Muster, Trends und Korrelationen, gewonnen werden können (aus der Berichtsperspektive: Data-Mining im engeren Sinn). Betrachtet man Data-Mining als Prozess im weiteren Sinn, wird deutlich, dass die Erhebung, Bereitstellung und Analyse von Daten sowie der Umgang mit den resultierenden Informationen oder algorithmischen Systemen weit über den Regulierungshorizont des Urheberrechts hinausgehen. Einige Aspekte wie der Datenschutz oder die Datennutzung in öffentlichen Aufgabenbereichen werden bereits seit Jahren reguliert, wobei Data-Mining meist unter dem Oberbegriff Datenverarbeitung subsumiert wird. Der Umgang mit daraus resultierenden Informationen, datenanalytischen Werkzeugen und digitalen Anwendungen wird erst ansatzweise in einigen Anwendungsbereichen reguliert.

Schutz und Nutzung von Daten

Der Schutz und die Nutzung (insbesondere komplexe Analysen) von Daten können über zwei Spezifika erschlossen werden. Zum einen haben Daten Kontextbezüge zu vielfältigen Sachverhalten, teilweise auch zu einzelnen Personen. Letztere sind besondere Datenobjekte, weil sie Grundrechte haben (u. a. auf informationelle Selbstbestimmung, Privatheit, Gleichbehandlung, Unversehrtheit, Eigentum und Informationszugang), die bei jeglicher Datenverarbeitung zu schützen sind. Zum anderen verfügen nur Unternehmen, Institutionen oder öffentliche Einrichtungen über die nötigen Ressourcen für komplexe Datenanalysen. Die Datenschutz-Grundverordnung[1] (DSGVO) zielt darauf ab, die Schutzinteressen betroffener Personen mit den Nutzungsinteressen datenverarbeitender Stellen abzuwägen. Grundsätzlich bedarf jegliche Verarbeitung personen-

1 Verordnung (EU) 2016/679 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung) (DSGVO)

beziehbarer Daten der freiwilligen und informierten Einwilligung durch Betroffene (Standard in privatwirtschaftlichen Bereichen) oder gesetzlicher Grundlagen (Standard bei Aufgaben im öffentlichen Interesse). Im zweiten Fall behalten nationale Regelungen zur Datenerhebung, zur primären Nutzung sowie zu Weiterverwendungsmöglichkeiten ihre Gültigkeit. Die DSGVO gilt nicht bei der Verarbeitung von Daten ohne Personenbezüge (z. B. vollständig anonymisierte Daten, Daten über Personengruppen, Geo- oder Umweltdaten, Metadaten).

Betroffene Personen haben gegenüber datenverarbeitenden Stellen *Rechte* auf Auskunft, Widerspruch, Berichtigung, Löschung, Übermittlung und Verarbeitungsbeschränkung sowie das Recht, keiner nur auf automatisierter Datenanalyse beruhenden rechtlich relevanten Entscheidung unterworfen zu werden. Außerdem haben sie Beschwerde- und Klagerechte. Datenverarbeitende Stellen können im Rahmen der ihnen vertraglich erteilten Nutzungsrechte personenbeziehbare Daten analysieren, müssen dabei jedoch *Grundsätze* (u. a. rechtmäßige, zweckgebundene, transparente und vertrauliche Verarbeitung) und *Pflichten* (zur Dokumentation, Rechenschaft, Einhaltung angemessener Sicherheitsmaßnahmen) einhalten. Die Spanne der Sicherheitsmaßnahmen reicht von Pseudonymisierung und Verschlüsselung der Daten über IT-Schutz vor Cyberattacken bis zur Datenschutzfolgenabschätzung bei Analyseprojekten, die erhöhte Risiken der Grundrechteverletzung bergen. Bei Verstößen drohen Geldbußen, bei finanziellen und gesundheitlichen Schäden haften datenverarbeitende Stellen.

Datenweiterverwendungen, u. a. zu wissenschaftlichen Forschungszwecken, gelten als mit einem ursprünglichen Erhebungszweck vereinbar. Diesbezüglich sollen Schutz- und Nutzungsinteressen spezifisch abgewogen und Einwilligungen möglichst eingeholt werden – eine Forderung, die nicht immer eingehalten werden kann (u. a. weil einige Pseudonymisierungsverfahren keine späteren Einwilligungen zulassen). Datentreuhandverfahren, bei denen neutrale Stellen Schutz- und Nutzungsinteressen spezifisch abwägen, werden in solchen Situationen oftmals als vermittelndes Bindeglied eingesetzt. Aus der Perspektive des Grundrechtsschutzes sollte dieses Bindeglied durch persönliche Einwilligungsmanagementsysteme ergänzt werden.

Datenverarbeitende Stellen haben als juristische Personen ihrerseits Rechte auf Geheimhaltung (Geschäfts- bzw. Amtsgeheimnisse). Unternehmen und Forschungseinrichtungen können ihre Tätigkeitsfelder im Rahmen des geltenden Rechts frei bestimmen und dürfen ihre Leistungen allein verwerten. Sie müssen bisher in der Regel weder Daten noch interne Data-Mining-Aktivitäten und deren Ergebnisse allgemein offenlegen. Zudem ist die Leistung zur Erstellung komplexer Datensätze (u. a. Bilder, Filme, Karten, Pläne) und zur Zusammenstellung großer Bestände (Datenbanken) urheberrechtlich durch alleinige Verwertungsrechte trotz Veröffentlichung geschützt – wobei Data-Mining zu wissenschaftlichen Forschungszwecken seit 2018 explizit zulässig ist. Datenbe-

reitstellungen können vertraglich vereinbart werden (z. B. bei öffentlicher Förderung). Das datenanalytische Aufgabenspektrum öffentlicher Einrichtungen ist gesetzlich definiert. Es gibt vielfältige bereichsbezogene gesetzliche Regelungen auf Bundes- oder Landesebene zu datenanalytischen Möglichkeiten und Pflichten sowie Weiterverwendungsmöglichkeiten und -grenzen. In der Summe entsteht bei Aufgaben im öffentlichen Interesse oftmals ein mehrschichtiges System von Erlaubnis- und Nutzungstatbeständen. Eine detaillierte Rechtsbetrachtung kann daher nur aufgabenbezogen erfolgen.

Datenverarbeitende Stellen können aufgrund gesetzlich oder vertraglich definierter Schutz- und Nutzungsrechte große Datenbestände aufbauen, exklusiv analysieren und verwerten und dadurch marktdominierende oder gar monopolartige Stellungen einnehmen, wodurch Wettbewerbsstrukturen verzerrt und Innovationen erschwert werden könnten. Um den Datenmonopolen in öffentlichen Aufgabenbereichen entgegenzuwirken, werden zum einen öffentliche Einrichtungen zunehmend verpflichtet, Daten unter Achtung von Schutzverpflichtungen zugänglich zu machen (Stichwort Open Data). Zum anderen werden (Forschungs-)Datenzentren und -infrastrukturen aufgebaut, über die Weiterverwendungen rechtssicher realisiert werden sollen. Das nationale Geoinformationswesen gilt als ein Vorreiter beim Aufbau von Dateninfrastrukturen und von Open-Data-Ansätzen bei Daten, die keinen Schutztatbeständen unterliegen. Zwar zielen diverse Gesetze auf Bundes- und Landesebene auf Verbesserungen des Zugangs zu Daten, die im Rahmen öffentlicher Aufgaben entstanden. Inwiefern damit eine stärkere Nutzung real gelingt, kann bisher kaum bewertet werden. Auf europäischer Ebene werden derzeit Verfahren diskutiert, um Datenmonopolen auch in privatwirtschaftlichen Strukturen entgegenzuwirken.

Rechtliche Herausforderungen in Bezug auf Data-Mining-Ergebnisse

Data-Mining-Prozesse können Informationen zu Strukturen in Datenbeständen sowie verallgemeinerbare Regeln und datentrainierte Modelle hervorbringen, die ggf. zu digitalen Anwendungen, algorithmischen Systemen oder allgemein zu Software weiterentwickelt werden können. Entsprechende wissenschaftliche Aktivitäten werden sowohl datenschutz- als auch urheberrechtlich privilegiert und im Rahmen der Forschungsförderung umfangreich unterstützt. Für die kontinuierliche Anwendung müssen diese Ansätze in der Regel jedoch in gemeinnützige oder gewerbliche Strukturen überführt und weiterentwickelt werden. Zudem wird Data-Mining auch gänzlich im Rahmen öffentlicher Aufgaben sowie gewerblicher Aktivitäten betrieben. Einrichtungsintern können Resultate als Geschäfts-, teilweise auch als Amtsgeheimnis deklariert werden.

Herstellerverantwortlichkeiten greifen – wenn überhaupt – erst, wenn Informationen oder Software für Dritte erstellt werden. Die Informationsgenerierung für Dritte wird rechtlich als *Dienstleistung* aufgefasst. Dienstleistungen werden vertraglich vereinbart (privatwirtschaftlicher Bereich) oder gesetzlich

definiert (öffentlicher Aufgabenbereich). Beide Wege bieten die Möglichkeit, Verantwortlichkeiten in Bezug auf Richtigkeits-, Sicherheits- und Haftungsfragen weitgehend auszuschließen. Algorithmische Systeme sind als Softwarebestandteile rechtlich *Produkte*. Sie werden vom allgemeinen Produktrecht erfasst, sofern nicht in einzelnen Anwendungsbereichen spezielle Regularien existieren. Zwar sind Hersteller während der gesamten Lebensdauer ihres Produktes für dessen Sicherheit verantwortlich, jedoch sind Hersteller nur in bestimmten Produktbereichen, die mit besonderen Risiken für Mensch und Umwelt einhergehen, verpflichtet, vor der allgemeinen Anwendung explizite Sicherheits- und Leistungsnachweise zu erbringen sowie Risiken und Nutzen einzelner Verfahren abzuschätzen und zu überwachen (z. B. in der Medizin).

Je nach Anwendungskontext und Aufgabenstellung können Data-Mining-Prozesse mit unterschiedlichen Folgedimensionen sowohl für einzelne Personen (auch solche, die keine Datengeber waren) als auch für die Gemeinschaft und die freiheitliche Grundordnung einhergehen. Da mögliche Folgen aus derartigen Aktivitäten weder mit den bisherigen Datenschutzregeln noch mit dem allgemeinen Produktrecht adäquat adressiert werden können, werden Forderungen nach einem eigenen Rechtsrahmen für datentrainierte algorithmische Systeme zur Entscheidungsfindung (teilweise kurz als KI-Systeme bezeichnet) lauter. Erste Konzepte sprechen sich für einen risikoadaptierten Regulierungsansatz aus, der unterschiedliche Schadensdimensionen minimieren und Rechtssicherheit in der Entwicklung und Anwendung algorithmischer Systeme bringen soll. Sie knüpfen an die etablierten Verfahren des Medizinprodukterechts an.

Meinungsumfragen zeigen regelmäßig, dass betroffene Personen trotz definierter Schutzelemente skeptisch sind gegenüber datenverarbeitenden Stellen und bezweifeln, eine ausreichende Kontrolle über die Verarbeitung ihrer Daten zu haben. Die Verbesserung der Grundrechtedurchsetzung ist eine der gesellschaftlichen Herausforderungen im Kontext der zunehmenden analytischen Möglichkeiten, die datenverarbeitende Stellen haben.

Diverse Fachgremien und Kommissionen setzen sich seit einigen Jahren mit den immer umfangreicheren Datenerhebungen, den kontinuierlich größer werdenden Datenbeständen, den zunehmenden Möglichkeiten der Verknüpfung und Analyse und den damit einhergehenden gesellschaftlichen Herausforderungen auseinander, um sicherzustellen, dass diese Möglichkeiten unter Wahrung der freiheitlichen Grundordnung und zum Wohle der Menschen realisiert werden. Die daraus resultierenden Empfehlungen von Enquete- und Ethikkommissionen sowie die Stellungnahmen zahlreicher Fachvertretungen gelten mehrheitlich auch in Bezug auf Data-Mining-Prozesse. Statt deren Empfehlungen auf der allgemeinen datenanalytischen Ebene (u. a. zur Digitalisierung zahlreicher Prozesse, zur Schaffung interoperabler Datenstrukturen, zum Ausbau nationaler Kompetenzen und der Weitentwicklung des Rechtsrahmens) zu reproduzieren, werden in der vertiefenden Betrachtung bereichsspezifische Herausfor-

derungen für Data-Mining in der Medizin und im Gesundheitssystem in den Blick genommen – zwei Bereiche, denen regelmäßig sowohl besondere Potenziale für den Einsatz komplexer datenanalytischer Verfahren als auch besondere Herausforderungen im Kontext der Digitalisierung zahlreicher Prozessabläufe unterstellt werden.

Data-Mining in der Medizin

Die Erhebung vielfältiger Daten und deren Analyse sind in der Medizin seit langem fest verankert, die Entwicklung und der Einsatz algorithmischer Systeme zur Behandlung von Krankheiten spezifisch reguliert.

Erhebung und Verwaltung medizinischer und administrativer Daten

Medizinische Daten werden aufgrund der Entwicklung unterschiedlicher Erhebungstechniken (z. B. bildgebende Verfahren, Sequenzierungstechniken, Tests) kontinuierlich detailgenauer. Sie bilden Kernbereiche der Persönlichkeit von Patient/innen ab und müssen in besonderem Maße geschützt und gesichert werden. Für Data-Mining vorrangig relevante medizinische Daten werden vor allem in Studien- oder Behandlungskontexten generiert. In *klinischen Studien* werden sehr viele Daten entsprechend der jeweiligen Studienpläne standardisiert erhoben, geprüft, befundet und analysiert (Rechtsgrundlage freiwillige informierte Einwilligung in die Nutzung zu Forschungs- und Entwicklungszwecken). Man unterscheidet öffentlich finanzierte Studien, deren Daten und Analysen der allgemeinen Wissenserweiterung dienen, von industriefinanzierten Studien, mit deren Daten die Sicherheit und Wirksamkeit/Leistungsfähigkeit neuer therapeutischer Produkte nachgewiesen werden. Die durch öffentlich finanzierte Studien generierten Daten werden zunehmend über Treuhandstrukturen verwaltet und auf Antrag bereitgestellt. Daten kommerzieller Studien können von den jeweiligen Sponsoren bisher allein genutzt werden. Zentrale Repositorien oder Dateninfrastrukturen für klinische Studiendaten gibt es bislang nicht.

Im Rahmen der *medizinischen Behandlung* entscheiden Ärzt/innen anhand ihrer fachlichen Expertise und möglicherweise existierender Behandlungsleitlinien situativ, welche patientenbezogenen Daten erhoben und befundet werden. Sie sind zur Dokumentation ihres Vorgehens und zur Speicherung relevanter Behandlungsdaten in dezentralen arztgeführten Primärakten gesetzlich verpflichtet. Diese Akteninhalte dienen der medizinischen Behandlung und der Klärung von Haftungsfragen. Für sekundäre Data-Mining-Aktivitäten sind sie bisher weder technisch noch rechtlich unmittelbar nutzbar. Viele Aktenbestandteile können zwar maschinell verwaltet, wegen der geringen Standardisierung bisher jedoch kaum maschinell analysiert werden. Zudem unterliegen sie stren-

gen Datenschutzvorgaben und der ärztlichen Schweigepflicht. Einrichtungsüberschreitende Datenzusammenführungen und -analysen bedürfen entweder weiterer gesetzlicher Regelungen (für Aufgaben im öffentlichen Interesse, wie z. B. die Überwachung von Krankheitsausbreitungen) oder der Einwilligung betroffener Patient/innen (u. a. um Teile von Behandlungsdaten in krankheitsspezifische Register oder medizinische Forschungsdatenbanken einstellen zu können). Bisher werden diese Einwilligungen zu möglichst weit formulierten Forschungszwecken (Broad Consent) schriftlich eingeholt und mit Treuhandstrukturen Datenzusammenführungen und Weiterverwendungen auf Antrag und nach Prüfung organisiert. Wenn die seit 2021 angebotene, von Patient/innen selbst zu führende, sekundäre elektronische Patientenakte in der Praxis auf allgemeine Akzeptanz trifft, sollen Patient/innen mit dieser ihre bisher verteilt gespeicherten Behandlungsdaten nach und nach zusammenführen sowie perspektivisch auch ihr Einwilligungsmanagement u. a. bezüglich weiterer Datenanalysen bis hin zur Datenspende zu Forschungszwecken darüber organisieren können. Ob sich die damit verbundenen datenanalytischen Hoffnungen realisieren lassen, bleibt abzuwarten.

Medizinische Einrichtungen müssen für diverse administrative Aufgaben standardisierte Datensätze zusammenstellen und an unterschiedliche Institutionen übermitteln. Für die Leistungsabrechnung extrahieren sie z. B. definierte patientenbezogene Diagnose- und Behandlungsinformationen aus den jeweiligen Primärakten, codieren sie anhand spezifischer administrativer Klassifikationen und leiten die jeweiligen Vergütungskennziffern bzw. -pauschalen ab. Aufgrund der vergleichsweise geringen Digitalisierung vielfältiger Prozessabläufe und der begrenzten Interoperabilität diverser IT-Systemkomponenten sind sowohl Behandlungsdokumentationen als auch Datenaufbereitungen, Codierungen und Datenzusammenstellungen personell aufwendig. Längst nicht alle in den Primärakten enthaltenen medizinischen Daten werden aufbereitet und codiert. Derartige Aufbereitungen sind notwendige Vorbereitungen für Data-Mining.

Medizinprodukte: Ausgangspunkt und Resultat von Data-Mining in der Medizin

Das Medizinprodukterecht rahmt Data-Mining-Prozesse in der Medizin in besonderem Maße: Zum einen sind die Messgeräte, die Daten zur Diagnose und zur Behandlung von Krankheiten erfassen, Medizinprodukte. Zum anderen fallen aus Data-Mining möglicherweise resultierende algorithmische Systeme, sofern sie zu digitalen Anwendungen oder Softwarebestandteilen weiterentwickelt werden und krankheitsbezogene Informationen über einzelne Personen liefern, unter das Medizinprodukterecht. Dieses Recht definiert ein spezifisches *Qualitätsmanagementsystem* mit unterschiedlichen Elementen je nach Gesund-

heitsschädigungspotenzial des Produktes. Ziel ist die Gewährleistung einer hohen Produktsicherheit, für die die jeweiligen Hersteller die Verantwortung tragen. Sie müssen die Sicherheit und Leistungsfähigkeit ihres Verfahrens prüfen (Produktentwicklungsphase), bevor es zertifiziert (Markteintritt) und umfangreich eingesetzt werden kann (Anwendungsphase).

Im Medizinprodukterecht werden vier Risikoklassen unterschieden. Zur niedrigsten Risikoklasse gehören Messgeräte, die lediglich krankheitsbezogene Vitalwerte aufzeichnen und anwendende Personen (Ärzt/innen oder Patient/innen) informieren. In die zweite Risikoklasse gehören einfache Berechnungsverfahren, mit denen u. a. Trends oder einfache medizinische Risikoscores (z. B. zur Bewertung von Erkrankungswahrscheinlichkeiten) berechnet werden. Anwendende Personen interpretieren diese datenbasierten Informationen selbst, sie ziehen Schlüsse, treffen Entscheidungen und bleiben verantwortlich. Medizinprodukte dieser beiden Klassen gelten insgesamt als relativ risikoarm. Hersteller prüfen eigenverantwortlich deren Sicherheit und Leistung und indizieren mittels CE-Kennzeichen, dass sie gesetzliche Vorgaben einhalten. Prüfinstanzen können Einsicht in die entsprechenden Unterlagen nehmen und weitere Prüfungen vornehmen. Zu hohen Risikoklassen gehören zum einen Verfahren, die Ärzt/innen bei Diagnose- und Therapieentscheidungen maßgeblich unterstützen und die bei Anwendungsfehlern gravierende Gesundheitsgefahren mit sich bringen (z. B. Monitoring von Herzfunktionen). Zum anderen sind es Verfahren, die Behandlungsentscheidungen automatisiert treffen und Ärzt/innen situativ ersetzen (z. B. Dosierung und Verabreichung hochwirksamer Arzneimittel). Bei Medizinprodukten hoher Risikoklassen müssen Hersteller sowohl Sicherheit und Leistung als auch den gesundheitsbezogenen Nutzen mit klinischen Studien nachweisen. Prüfinstanzen nehmen die entsprechende Zertifizierung vor.

Auch in der Anwendungsphase muss die Qualität jedes Medizinprodukts gesichert werden, u. a. durch Produktregistrierungen, genaue Anwendungsinstruktionen (z. B. Einsatz nur durch Fachkräfte), regelmäßige technische Verfahrensprüfungen, ein kontinuierliches Risikomonitoringsystem während der Anwendung (Vigilanz) oder Verwendungskorrekturen bis hin zu Rückrufen. Die Umsetzung der Qualitätssicherungsmaßnahmen ist vor allem bei Medizinprodukten hoher Risikoklassen aufwendig.

Die Maßnahmen zur Sicherung einer hohen Produktqualität werden ergänzt durch Haftungsregeln: Hersteller haften bei Schäden sowohl in der Produktentwicklungsphase (Gefährdungshaftung) als auch in der Anwendungsphase (Verschuldenshaftung). Um Anwendungsrisiken zu reduzieren, werden risikoreichere Medizinprodukte meist in ärztlicher Verantwortung verwendet. Damit teilen sich Medizinproduktehersteller und anwendende Ärzt/innen die Verantwortung. Zugleich werden Schuldnachweise komplexer, die über Sorgfaltspflichtverletzungen von Herstellern oder Ärzt/innen nachgewiesen werden müssen. Beim Einsatz datentrainierter kontinuierlich lernender Systeme dürfte es für Ge-

schädigte noch schwerer werden, ein Verschulden von Herstellern und anwendenden Ärzt/innen nachzuweisen, denn auch die Trainingsdaten können die Qualität des Medizinproduktes beeinflussen, Ursache für Fehleinschätzungen sein und dadurch Gesundheitsschäden verursachen, ohne dass Hersteller oder Ärzt/innen ihre Sorgfaltspflichten verletzten.

Die Weiterentwicklung des medizinischen Haftungsrechts in Bezug auf datentrainierte algorithmische Systeme zur Entscheidungsfindung ist somit eine sehr wichtige und große Herausforderung.

Integration datentrainierter Algorithmen in die medizinische Versorgung

Mit der Zertifizierung sind Medizinprodukte zwar marktverfügbar, für die breite Anwendung ist jedoch die Akzeptanz des Einsatzes sowohl bei medizinischen Fachgesellschaften (Aufnahme in Behandlungsleitlinien) und Kostenträgern (Aufnahme in die Leistungskataloge der Krankenkassen) als auch bei Patient/-innen und behandelnden Ärzt/innen wichtig. Dafür spielen neben der Praktikabilität während der Behandlung vor allem Fragen zum mit dem Medizinprodukteeinsatz verbundenen (Zusatz-)Nutzen eine entscheidende Rolle. Dieser Nutzen kann auf unterschiedliche Art und Weise definiert und bewertet werden. Gesetzliche Krankenkassen fordern zunehmend Belege für einen gesundheitsbezogenen Nutzen, bevor sie die Kosten für den Einsatz entsprechender Verfahren im Rahmen der Regelversorgung tragen. Zwar wurde die Vorgehensweise für die Aufnahme medizinischer Apps bzw. Medizinprodukte der unteren zwei Risikoklassen in die Leistungskataloge der gesetzlichen Krankenkassen 2020 beschleunigt. Dennoch ist das Verfahren aufwendig im Vergleich zu nichtmedizinischen Produkten, wie z. B. Gesundheits-Apps. Der Aufwand für die Integration datenbasierter algorithmischer Systeme in die medizinische Versorgung steigt mit der Risikoklasse, in die sie eingestuft werden. Die für klassische Medizinprodukte existierenden vielschichtigen Wege in die medizinische Versorgung sollten bezüglich ihrer Passfähigkeit für datenbasiert lernende Entscheidungs(unterstützungs)systeme geprüft und ggf. weiterentwickelt werden. Dies erfordert situative Betrachtungen einzelner Verfahren in spezifische Anwendungskontexten.

Anwendungsbeispiel: algorithmenbasierte Mammografiebefundung

Seit Jahren wird die Mammografiebefundung als ein möglicher Einsatzbereich für datentrainierte Algorithmen zur Bilderkennung genannt, u. a. weil im Rahmen nationaler Screeningprogramme Mammografieaufnahmen massenhaft erzeugt, fachärztlich befundet und Befundergebnisse durch weitere Untersuchungen verifiziert und geprüft werden. Um in der medizinischen Praxis bestehen zu können, müssen zum einen Sicherheit und Leistungsfähigkeit dieser datentrai-

nierten Algorithmen geprüft und diese zertifiziert sein. Zum anderen müssen medizinische Fachgesellschaften und Kostenträger einen (Zusatz-)Nutzen zum Status quo der Mammografiebefundung anerkennen. Mammografiescreeningprogramme werden national eigenständig organisiert. In etlichen europäischen Ländern befunden zwei Radiolog/innen unabhängig voneinander, bei Unstimmigkeiten wird zu dritt beraten, bei Auffälligkeiten folgen weitere Diagnose- und Behandlungsschritte. Im US-amerikanischen Programm befundet auf der ersten Stufe nur ein Radiologe bzw. eine Radiologin. Ziel der Programme ist die frühe Erkennung von Brustkrebs. Befürworter/innen betonen, dass dadurch schwere Behandlungsverläufe reduziert werden. Kritiker/innen verweisen auf die Überdiagnostik, die in Überbehandlung münde und mehr schade als nütze.

Die erste Generation datentrainierter Assistenzsysteme zur Mammografiebefundung baute auf symbolischen Verfahren auf, erzielte in den USA unter Studienbedingungen gute Ergebnisse, wurde zertifiziert und dort in kurzer Zeit nahezu flächendeckend eingesetzt. In der Routineanwendung stellte sich heraus, dass durch den Einsatz dieser Assistenzsysteme weder die Befundqualität gehalten (Überdiagnosen nahmen leicht zu) noch der erhoffte Zeitgewinn realisiert werden konnten. Die neueste Generation solcher Assistenzsysteme nutzt datentrainierte künstliche neuronale Netze und konnte unter Studienbedingungen im US-amerikanischen Programm erstmals leichte Verbesserungen zur Einzelbefundung erzielen. Im europäischen Ansatz mit Doppelbefundung und Konsensusentscheidung wurden vergleichbare Ergebnisse, aber keine Verbesserungen erzielt.

Befürworter/innen derartiger Assistenzsysteme betonen, dass vor allem datentrainierte künstliche neuronale Netze in der medizinischen Diagnostik zunehmend in Anwendungsnähe kommen (z. B. bei der Erkennung von Hautkrebs oder Herzrhythmusstörungen) und fordern mehr nationales Engagement. Trainings- und Testdaten, die die nationale Bevölkerung repräsentativ abbilden, sollten erstellt, Zertifizierungsverfahren und spezifische Qualitätssicherungssysteme definiert werden. Zudem sollte die Akzeptanz bei den Akteuren der jeweiligen medizinischen Versorgungstrukturen in den Blick genommen werden. Bezüglich der algorithmenbasierten Mammografiebefundung sind die relevanten nationalen Fachgesellschaften bisher eher skeptisch. Sie halten die Umstellung von der 2-D- auf die 3-D-Aufnahmetechnologie für vielversprechender als die automatisierte Bildbefundung. Wenn sich die Aufnahmetechnologie ändert, dann müssen auch neue Trainingsdaten erstellt sowie Algorithmen erneut trainiert, geprüft und zertifiziert werden.

Die Technikdiffusion ist bei allen risikoreichen Medizinprodukten eine Herausforderung, egal ob sie aus Data-Mining-Prozessen resultieren oder nicht. Sinnvollerweise könnte daher zunächst ein Verfahren als Add-on-Technologie in speziellen Zentren unter Alltagsbedingungen weiter getestet werden. Zeichnet sich in der Anwendung ein Zusatznutzen im Vergleich zu den etablierten

Verfahren ab, kann eine Integration in die Regelversorgung abgestimmt werden. In der Folge könnten sich u. a. Fragen zur Veränderung organisatorischer Arbeitsabläufe ergeben, die personalintensive Doppelbefundungen könnten zur Disposition gestellt und Ressourcen umverteilt werden. Fraglich ist, ob durch derartige Assistenzsysteme Kosten gesenkt und Patient/innen mit schweren Erkrankungen in der Praxis intensiver betreut werden könnten, oder ob nur der Dokumentationsaufwand weiter steigt.

Datenstrukturen und Data-Mining im Gesundheitssystem

Seit Jahren werden diverse gesundheitssystemische Aufgaben mithilfe strukturerkennender datenanalytischer Verfahren realisiert. Die Spanne reicht von der Fehlersuche in administrativen Daten bis zur Fortschreibung von Gruppierungsalgorithmen zur Klassifikation erbrachter Behandlungsleistungen für die Fallpauschalenzuweisung oder versicherter Personen für den morbiditätsorientierten Risikostrukturausgleich gesetzlicher Krankenkassen. Zudem werden in vielfältigen Forschungsprojekten Strukturen und Muster in verfügbaren Datenbeständen gesucht, um z. B. Informationen zur räumlichen Verteilung oder zeitlichen Entwicklung von Gesundheitsrisiken zu gewinnen, nach unerwünschten gesundheitlichen Ereignissen durch therapeutische Maßnahmen zu suchen oder die Qualität medizinischer Leistungen abschätzen zu können.

Die derzeit wichtigste Basis für solche Data-Mining-Prozesse sind die gesetzlich definierten Leistungsabrechnungsdaten, die medizinische Einrichtungen erstellen und an unterschiedliche Institutionen des nationalen Gesundheitssystem hochgradig standardisiert übermitteln. Diese Datensätze haben vielfältige Bezüge und Informationen: zu Patient/innen und deren gesundheitlicher Situation, zu Ärzt/innen, medizinischen Einrichtungen und deren Behandlungsleistungen, zu Krankenkassen und deren Leistungsvergütung sowie zu Zeit und Raum. Alle Personen und Einrichtungen sind über eindeutige Nummern lebenslang identifizierbar. Die Nummern dienen einerseits als Pseudonym und erlauben andererseits eine regelmäßige zeitliche Fortschreibung der Abrechnungsdatenbestände bei datenempfangenden Institutionen. Alle gesundheits-, behandlungs- und vergütungsbezogenen Angaben werden mit Klassifikationen codiert, die für administrative Zwecke entwickelt wurden (es sind keine medizinisch hochdifferenzierenden Nomenklaturen). Die Datenzusammenstellung ist für medizinische Einrichtungen aufwendig, denn Diagnosen und Behandlungsleistungen werden nicht automatisiert codiert. Oftmals gibt es vergütungsrelevante Ermessensspielräume. Umfangreiche Datenprüfungen sind erforderlich, dennoch können Fehler und Verzerrungen (z. B. Überdiagnosen, Doppelerfassungen) nicht gänzlich ausgeschlossen werden. In der Summe bilden diese maschinell gut verarbeitbaren Versorgungsdaten sowohl die gesundheitliche Situation gesetzlich Versicherter als auch Behandlungs- und Abrechnungsprozesse auf

Einzelfallebene im Zeitverlauf vollständig ab. Auch wenn sie keine allzu hohe medizinische Detailgenauigkeit aufweisen, haben sie gleichwohl ein erhebliches analytisches Potenzial. Ein komplexes Regelwerk definiert für jede Institution der gesundheitssystemischen Selbstverwaltung

- welche Datenbestände sie aufbauen darf,
- welche datenanalytischen Aufgaben sie in welchem Umfang eigenverantwortlich realisieren soll,
- welche Daten sie ggf. an wen weiterleiten muss sowie
- welche Datenschutzkonzepte jeweils gelten.

Neben diesen gesetzlich definierten primären datenanalytischen Aufgaben (im öffentlichen Interesse) haben einzelne Institutionen zudem sekundäre Analysemöglichkeiten (z. B. für Entwicklungs- oder Forschungsaktivitäten), bei denen ein Kontrollgremium situativ Schutz- und Nutzungsinteressen bei komplexen Datenanalysen bzw. Data-Mining-Prozessen abwägt. Mitunter dürfen die Institutionen dafür spezielle datenanalytische Abteilungen einrichten oder Institutionen gründen.

Zwar legitimieren die primären Aufgaben im öffentlichen Interesse die Beschränkungen der informationellen Selbstbestimmung. Das Fehlen jeglicher Widerspruchsmöglichkeiten für Betroffene in sekundäre Weiterverwendungen derartiger Daten besonderer Kategorie wird mitunter jedoch als paternalistische Fremdbestimmung kritisiert, zumal unterschiedliche Teilbestände in mehreren Etappen weitergeleitet und von unterschiedlichen Institutionen weiterverwendet werden können. Beispielsweise fließen personenbezogene ambulante Rezeptdaten zeitnah zur Arzneimittelabgabe bei wenigen deutschlandweit agierenden Apothekenrechenzentren zusammen, die diese Daten einerseits zur Abrechnung mit den unterschiedlichen Krankenkassen benötigen, denen sie die Rezeptdaten ihrer Versicherten dazu weiterleiten. Andererseits können die Rechenzentren anonymisierte Rezeptdaten zeitnah selbst weiterverwenden, um z. B. datenbasierte Informationsdienste für ihre Kund/innen zu entwickeln oder das Marktgeschehen zu analysieren. Sie können diese Daten auch anderweitig verwerten und z. B. Nutzungsrechte verkaufen.

Auch Krankenkassen können Rezeptdaten ihrer Versicherten im Rahmen ihrer Aufgaben und zu Forschungszwecken relativ zeitnah selbst verwenden, allerdings nicht anderweitig verwerten. Zudem übermitteln alle Krankenkassen pseudonymisierte Jahresdatensätze (mit allen Leistungsabrechnungsdaten ihrer Versicherten) an den Spitzenverband der Krankenkassen. Dieser prüft alle Jahresdatensätze, nutzt sie im Rahmen seiner Aufgaben selbst und leitet sie weiter an das Bundesamt für soziale Sicherung (zur Fortschreibung des morbiditätsorientierten Risikostrukturausgleichs) und an das inzwischen beim Bundesinstitut für Arzneimittel und Medizinprodukte angesiedelten Forschungsdatenzentrum, das alle Versorgungsdaten speichert, kontinuierlich fortschreibt und in en-

gen Grenzen für nichtkommerzielle Forschungszwecke bereitstellt. Bisher war das Nutzungsinteresse an den vom Forschungsdatenzentrum bereitgestellten Daten begrenzt. Ein Grund sind die bisherigen mehrjährigen Zeitverzögerungen bis zur Bereitstellung, ein anderer die engen Grenzen der Nutzungsberechtigung. Ausschließlich Institutionen der gesundheitssystemischen Selbstverwaltung und der wissenschaftlichen Forschung sind antragsberechtigt. Einige von ihnen (z. B. Krankenkassen) können jedoch aktuellere Teildatenbestände unmittelbar nutzen und dafür teilweise mit akademischen Institutionen kooperieren.

Gesundheitssystemische Data-Mining-Prozesse starten in der Regel zunächst in kleinerem Rahmen als Forschungsprojekte oder als Machbarkeitsstudien, deren Ergebnisse anschließend fachlich diskutiert werden. Dabei wird regelmäßig deutlich, dass auch methodisch und analytisch geeignete Verfahren nur solche Strukturen und Informationen extrahieren können, die in den Analysedatensätze enthalten sind. Eine räumlich zu geringe Auflösung kann keine lokalen Spezifika aufzeigen, alte Analysedaten können keine Risiken neuer Arzneimittel oder Behandlungsmethoden zeitnah sichtbar machen, und die Richtigkeit einzelner Angaben kann kaum rückwirkend geprüft werden. Jeder datenanalytische Ansatz und die resultierenden Ergebnisse werden in Fachkreisen diskutiert, situativ abgewogen und bewertet. Danach können Verfahren ggf. verstetigt und Algorithmen z. B. in epidemiologische Informationsdienste oder in größere gesundheitssystemische Prozesse, wie das Fallpauschalensystem, integriert werden. Einen Produktstatus erreichen dieserart Algorithmen in der Regel nicht.

Fazit und Handlungsoptionen

Data-Mining ist ein unscharfer Begriff – ähnlich wie Big- oder Smart Data, maschinelles Lernen oder künstliche Intelligenz. Sie alle werden mit komplexen datenanalytischen Verfahren assoziiert und haben erhebliche Schnittmengen, insbesondere wenn man die damit einhergehenden gesellschaftlichen Herausforderungen in den Blick nimmt. Mit diesen Begriffen verbundene Visionen beruhen oftmals auf der Grundannahme, dass immer mehr Daten die Realität so umfangreich abbilden, dass Regeln und Modelle weitgehende Allgemeingültigkeit erreichen und zur Klassifikation, Gruppierung oder Prognose neuer Sachverhalte eingesetzt werden können. Datenanalyst/innen betonen zudem, dass einerseits auch große Datenmengen reale Sachverhalte kaum vollständig abbilden und Regeln und Modelle stets Vereinfachungen einer komplexeren Realität seien, und dass andererseits real existierende strukturelle Probleme, wie z. B. die Diskriminierung einzelner Personengruppen, durch derartige Ansätze reproduziert werden könnten. Folglich können die Resultate derartiger Prozesse situativ nützlich sein, aber auch Risiken mit sich bringen. Die Schaffung eines Mehrwerts unter Achtung der freiheitlichen Grundordnung ist folglich eine Frage der Ausgestaltung derartiger Prozesse.

Zahlreiche Sachverständigenräte und Kommissionen auch des Deutschen Bundestags und der Bundesregierung haben sich in den letzten Jahren mit den Möglichkeiten und Herausforderungen der Digitalisierung im Allgemeinen sowie den wachsenden Datenbeständen, mit den Möglichkeiten und Grenzen deren Analyse und mit dem Umgang der Ergebnisse im Besonderen auseinandergesetzt sowie diesbezüglich Empfehlungen und Handlungsoptionen erarbeitet, zu denen wiederum zahlreiche Stakeholder Stellung genommen haben. Unisono wird empfohlen, Digitalisierungsaktivitäten zu forcieren, Infrastrukturen zur Weiterverwendung von Datenbeständen auf- und auszubauen, die Datennutzung stärker in den Blick zu nehmen, datenanalytisches Know-how zu stärken, die Entwicklung entsprechender Anwendungen zu fördern und risikoreiche stärker zu regulieren sowie eine größere nationale oder europäische digitale Souveränität anzuvisieren, auch um hohe Schutzstandards und die Grundrechtesicherung zu gewährleisten. Diese Empfehlungen lassen sich auch aus den Ausführungen dieses Berichts und den dafür in Auftrag gegebenen Gutachten ableiten. Bei einer vergleichenden Betrachtung unterschiedlicher datenanalytischer Anwendungsbereiche wird zudem deutlich, dass es bereichsspezifische Besonderheiten, Stärken und Schwerpunktsetzungen gibt, die sich teilweise zu ergänzen scheinen. Eine abschließende vergleichende Gesamtschau soll Handlungsoptionen für das Parlament fundieren.

Standardisierung, Zugänglichkeit und Nutzbarkeit von Daten verbessern

Der Geodatenbereich gilt als ein Vorreiter beim Aufbau interoperabler Dateninfrastrukturen über die standardisierte amtliche Referenzdatenbestände rechtssicher bereitgestellt werden. Die ursprüngliche Differenzierung der Datenbereitstellung für öffentliche Aufgaben, Forschungstätigkeiten und kommerzielle Weiterverwendungsabsichten wird zunehmend aufgegeben, Open-Data-Ansätze gewinnen an Bedeutung und Zugangshürden sinken. Wie intensiv die bereitgestellten Geodaten für komplexe Datenanalysen mittels Data-Mining tatsächlich genutzt werden, lässt sich bisher kaum abschätzen. Der Deutsche Bundestag könnte in den regelmäßig vorzulegenden Geo-Fortschrittsberichten Untersuchungen zur Entwicklung der Datennutzung einfordern, um die Potenziale der Datenangebote gezielter erfassen, bewerten und auszuschöpfen zu können.

Einrichtungen des Gesundheitssystems wird seit Jahren erheblicher Entwicklungsbedarf bezüglich der Digitalisierung unterschiedlicher datenverarbeitender Prozesse, der Entwicklung und Nutzung von Datenstandards und dem Aufbau interoperabler Datenzugangsstrukturen attestiert. Um den zukünftigen Aufwand für unterschiedliche Datenaufbereitungen zu senken, sollte die Entwicklung und Verwendung einheitlicher medizinischer Terminologien und interoperabler Formate bereits bei der primären Behandlungsdokumentation vorangetrieben und perspektivisch vorgeschrieben werden. Dabei gilt es den Arbeitsaufwand von be-

handelnden Ärzt/innen im Blick zu behalten und nach Lösungen zu suchen, die den Dokumentationsaufwand so gering wie möglich halten.

Die Daten der arztgeführten Primärakten unterliegen der Schweigepflicht und höchsten Datenschutzvorgaben. Sie werden in spezifischen Informationssystemen einrichtungsinternen gespeichert. Diese Systeme sind nicht für Data-Mining-Aktivitäten konzipiert. Um Behandlungsdaten dafür weiterverwenden zu können, müssen diese aufbereitet und in zumeist einrichtungsübergreifende sekundäre Register oder Datenzentren überführt werden. Dafür sind gesetzliche Regelungen (bei Aufgaben im öffentlichen Interesse) oder Einwilligungen erforderlich. Beide Verfahrensformen werden seit langem genutzt, um vielfältige, spezifisch definierte Datensätze aus den Primärakten abzuleiten und an unterschiedliche medizinische Register oder Datenzentren zu übermitteln, die diese Daten für administrative und gesundheitssystemische Aufgaben aber auch zu Forschungs- und Planungszwecken bereitstellen. Diese Register und Datenzentren fungieren als Datentreuhänder in vielfältigen Organisationsformen. Die bereits etablierten Datentreuhandformen sollten bezüglich ihrer Praktikabilität geprüft, weiterentwickelt und harmonisiert werden. Sie könnten beispielgebend auch für andere Bereiche sein, in denen geschützte Daten nicht monopolisiert gehalten, sondern unter Berücksichtigung relevanter ethischer Aspekte weiterverwendet werden sollen (z. B. Mobilitätsdaten).

Im Laufe der Zeit haben vielfältige spezialgesetzliche Regelungen zum Umgang mit gesundheitsbezogenen Daten in den unterschiedlichen Einrichtungen des Gesundheitssystems eine erhebliche Komplexität erreicht, die zu Unsicherheiten bezüglich der Möglichkeiten und Grenzen der Datenweiterverwendung führt und dadurch Datenanalysen erschwert. Bisher gibt es keine Gesundheitsdateninfrastruktur, die unterschiedliche Datenzentren und Register vernetzt und die Daten des nationalen Gesundheitswesens rechtssicher zugänglich macht. Das 2021 verabschiedete Datennutzungsgesetz[2], das die Nutzungsmöglichkeiten der Daten des öffentlichen Sektors harmonisieren und befördern soll, gilt nicht für die Daten, die im nationalen Gesundheitswesen verarbeitet werden. Ein diesbezügliches Spezialgesetz, das die Vielfalt der gesundheitssystemischen datenbezogenen Regelungen harmonisiert und vereinfacht, erscheint dringend geboten.

Mit der seit 2021 allen Versicherten anzubietenden elektronischen Patientenakte werden derzeit große Hoffnungen verbunden, vielfältige gesundheitsbezogene Daten vor allem aus Behandlungskontexten in der Verantwortung einzelner Patient/innen zusammenzuführen und perspektivisch auch das Einwilligungsmanagement für die Datenweitergabe bis hin zu Datenspenden zu Forschungszwecken damit zu organisieren. Dieses Einwilligungsmanagement ist

2 Gesetz für die Nutzung von Daten des öffentlichen Sektors (Datennutzungsgesetz – DNG)

von zentraler Bedeutung für sekundäre Datenverwendungen einschließlich Data-Mining. Wie viele Versicherte dieses Angebot zur Datenselbstverwaltung annehmen und in die Datenweiterverwendung zu Forschungszwecken einwilligen werden, ist derzeit unklar. Eine Begleitforschung Entwicklung zur Akzeptanz dieser Akten und der Nutzung unterschiedlicher Funktionalitäten und Services scheint dringend geboten. Das Parlament könnte sich berichten lassen.

Konkretion der privilegierten Datenverwendung zu Forschungszwecken

Datenweiterverwendungen zu Forschungszwecken einschließlich Data-Mining werden zum einen datenschutzrechtlich privilegiert, zum anderen begrenzen sie Urheber- bzw. Leistungsschutzrechte. Etliche Formulierungen zum Forschungsprivileg sind jedoch auslegungswürdig. Die DSGVO empfiehlt lediglich, den wissenschaftlichen Forschungsbegriff mit der Einhaltung anerkannter ethischer Forschungsstandards zu verknüpfen, ein entsprechendes Einwilligungsmanagement vorzusehen, Forschungsabsichten im Einzelfall zu prüfen und sowohl öffentliche als auch privatwirtschaftlich finanzierte Forschung bis hin zu technologischen Entwicklungen und Anwendungsdemonstrationen zuzulassen. Über Öffnungsklauseln lässt sie jedoch nationale Spezifikationen zu.

Ethische (Forschungs-)Standards und die Prüfung von Analyseanträgen sind in der Medizin und im Gesundheitssystem seit langem verankert. Handlungsbedarf gibt es derzeit vor allem bezüglich der Vereinheitlichung, Beschleunigung und Straffung der Antragsprüfungen. Das Einwilligungsmanagement in Datenweiterverwendungen ist derzeit bei medizinischen und gesundheitssystemischen Einrichtungen jedoch eine gewisse Schwachstelle, weil Einwilligungen bisher schriftlich erteilt werden müssen und rückwirkend oftmals nicht eingeholt werden können. Auch deshalb wird der Forschungsbegriff im deutschen Gesundheitssystem bisher enger ausgelegt als in der DSGVO. Für die in unterschiedlichen Registern und Datenzentren gespeicherten personenbezogenen Gesundheitsdaten sind in der Regel nur bestimmte öffentliche (Forschungs-)Einrichtungen nutzungsberechtigt, Forschungsabsichten müssen im öffentlichen Interesse liegen. Dadurch können u. a. Medizinproduktehersteller nur in Kooperation mit öffentlichen Forschungseinrichtungen entsprechende Daten nutzen, um z. B. algorithmische Assistenzsysteme zu trainieren.

Parallel dazu sind Unternehmen, die klinische Studien zum Sicherheits- und Leistungsnachweis von medizinischen Produkten finanzieren, nicht zur Bereitstellung ihrer Studiendaten verpflichtet. Vertreter/innen der freien Wirtschaft, der (medizinischen) Forschung sowie öffentlicher Einrichtungen kritisieren seit Jahren die derzeitigen Verfahren sowie die damit einhergehenden Ungleichbehandlungen und betonen im medizinischen Kontext die gesundheitsbezogenen Risiken durch die Nichtnutzung von Daten, wenn beispielsweise Erkrankungs-

risiken, Infektionsherde oder unerwünschte Nebenwirkungen von Behandlungsverfahren nicht erkannt werden. Die Etablierung offenerer Datennutzungskonzepte sollte daher diskutiert bzw. geprüft werden. Dazu könnten die Reichweite des Forschungsbegriffs und bestehende Datenverarbeitungsprivilegien diskutiert und gesetzlich klargestellt werden.

Qualitätsmanagementsysteme bei Medizinprodukten – Vorbild für den Umgang mit Data-Mining-Ergebnissen in anderen Bereichen?

Inwiefern Data-Mining-Prozesse zum Wohle der Gemeinschaft beitragen, Grundrechte Einzelner schützen oder gefährden, transparent gestaltet oder aber mit menschlichen Kontrollverlusten in Entscheidungssituationen einhergehen und welche Folgen daraus erwachsen, kann nur situativ abgewogen und bewertet werden. Die im medizinischen Kontext über Jahrzehnte entstandenen Verfahren zur Qualitätssicherung medizinischer Produkte mit ihren risikoajustierten abgestuften Zertifizierungsverfahren in Kombination mit kontinuierlichen produktbezogenen Sicherheitsprüfungen und Risikoüberwachungen während der Anwendung könnten beispielgebend für andere risikoreiche Anwendungsbereiche sein, in denen datenanalytische Verfahren und algorithmenbasierte Systeme zunehmend eingesetzt werden (z. B. innere oder äußere Sicherheits-, Fin- oder Legal-Tech-Bereiche). Die Forderungen nach risikoadaptierten Regulierungen und Algorithmen-TÜVs oder der derzeit auf europäischer Ebene verhandelte Digital Service Act greifen unterschiedliche qualitätssichernde Maßnahmen des Medizinprodukterechts bereits auf. Mit diesbezüglichen Vorgehensweisen, deren Konkretisierung und Harmonisierung vor allem in risikoreichen Anwendungskontexten sollten Stakeholder sich intensiver befassen. Dadurch könnten Analyst/innen und Prüfinstanzen wichtige Informationen zur Sicherheits- und Leistungsbewertung erhalten sowie Prüf- und Monitoringverfahren etabliert werden, mit denen Risiken während der Anwendung algorithmischer Systeme überwacht und ggf. reduziert werden könnten.

Die unterschiedlichen Elemente der im medizinischen Bereich etablierten Qualitätsmanagementsysteme zielen primär auf eine hohe Produktsicherheit und die Generierung eines gesundheitsbezogenen Nutzens durch die Produktanwendung ab. Jedoch lassen sich auch mit höchst umfangreichen Qualitätsmanagementsystemen beim Einsatz datentrainierter algorithmischer Systeme nie alle vorrangig gesundheitsbezogenen Risiken vollständig ausschließen, denn auch große Datenbestände und komplexe mathematisch-statistische Modelle bilden die Realität vereinfacht ab, kommen bei höchstseltenen Situationen an ihre Grenzen, können real existierende Diskriminierungen reproduzieren und liefern Ergebnisse, die mitunter selbst für Expert/innen im Detail nur schwer nachzuvollziehen sind. Deshalb sind die Klärung von dauerhaften Produktverantwortlichkeiten und von Haftungsfragen relevante Aspekte für die Akzeptanz und den Einsatz algorithmischer Assistenzsysteme. Forschungseinrichtungen,

die Daten privilegiert nutzen dürfen, um Modelle zu trainieren und Assistenzsysteme zu entwickeln, kommen regelmäßig bereits bei der Produktzertifizierung an ihre Grenzen. Die kontinuierliche Gewährleistung einer hohen Produktsicherheit und Haftung im Schadenfall gehört nicht mehr in das Tätigkeitsspektrum von Forschungseinrichtungen. Spätestens dafür sind wirtschaftlich agierende Unternehmen erforderlich. Bereits bei klassischen Softwareprodukten wird die Eignung des derzeitigen Haftungsrechts in medizinischen, aber auch in anderen Einsatzbereichen kontrovers diskutiert. Besondere haftungsrechtliche Herausforderungen ergeben sich durch kontinuierlich lernende, medizinische Assistenzsysteme. Produktverantwortlichkeiten und Haftungsfragen sollten daher systematisch und spezifisch durchdacht, abgewogen und rechtlich geklärt werden.

1 Einleitung

Hintergrund

Seit Jahren wird eine gewisse Dateneuphorie geschürt, die sich wesentlich darauf stützt, dass durch die Digitalisierung nahezu aller Lebensbereiche kontinuierlich nicht nur mehr Daten erhoben und gespeichert, sondern durch deren Analyse auch Informationen gewonnen, Erkenntnisse abgeleitet, Wissen erweitert und Nutzen gestiftet werden können. Dadurch sollen sich vielfältige Prozessabläufe verbessern, Effizienzgewinne erzielen sowie neue Informations-, Leistungs- oder Serviceangebote entwickeln lassen. Neue Begriffe mit semantischen Unschärfen lassen Spielraum zur Interpretation: Daten werden als Rohstoff der Wissensgesellschaft oder als Öl des 21. Jahrhunderts bezeichnet, wobei die Bestände immer größer (Big Data) und die Techniken zu deren Erhebung und Verwendung als immer smarter oder intelligenter bezeichnet werden. Dazu passt der in den 1990er Jahren eingeführte Begriff Data-Mining (wörtlich Datenbergbau), der mit dem suggestiven Bild des Schürfens nach Rohstoffadern und des Findens von Nuggets in den wachsenden Datenbeständen des heutigen Datenzeitalters spielt (Schepers et al. 2015, S. 20).

Damit einher geht jedoch auch eine diffuse Skepsis gegenüber ausufernder Erfassung, intransparenter Weiterleitung und Zusammenführung von Daten sowie immer ausgefeilteren Analysetechniken, insbesondere dann, wenn es um persönliche Daten geht. Diese Skepsis wird u. a. dadurch befördert, dass etliche Personen sich weder als Herrschende über ihre Daten sehen noch die Analysetechniken nachvollziehen und deren Intention abschätzen können.

Seit Jahren befassen sich Teile der Wissenschaft und Wirtschaft, der Zivilgesellschaft und der Politik intensiv mit den zunehmenden Möglichkeiten der Verknüpfung und Analyse vielfältiger Datenbestände, den damit verbundenen Innovationspotenzialen und gesellschaftlichen Herausforderungen sowie mit ethischen Aspekten und Regulierungsoptionen. Der Bundestag hat dazu Enquetekommissionen eingerichtet, sich mit vielfältigen Facetten der Thematik beschäftigt und umfangreiche Berichte verabschiedet; die Bundesregierung hat Ethikkommissionen beauftragt und zahlreiche Strategien zur Digitalisierung, dem Umgang mit Daten und künstlicher Intelligenz (KI) im Allgemeinen oder zu Geoinformationen oder Innovationen in der Medizintechnik im Besonderen entwickelt. Auch für die parlamentarische Technikfolgen-Abschätzung bilden vielfältige Fragen in Bezug auf die immer detailliertere Erfassung und umfangreichere Verknüpfung von Daten, die Entwicklung komplexer Analysetechniken bis hin zur gesellschaftsverträglichen Ausgestaltung digitaler Geschäftsprozesse seit einiger Zeit – und aller Voraussicht nach auch in den kommenden Jahren – einen Hauptuntersuchungsgegenstand. Dabei besteht eine zentrale

Herausforderung darin, die schnellen analysetechnischen Veränderungen sowie die sozioökonomischen und gesellschaftspolitischen Entwicklungen so zu analysieren, dass die Ergebnisse nicht bereits nach kurzer Zeit überholt erscheinen. Um das zu gewährleisten, müssen die Debatten zeitnah aufgegriffen und mit der derzeitigen realen Situation sowie substanziell begründeten Wissensbeständen, die u. a. aus früheren TA-Projekten resultieren, abgeglichen werden. Denn es sollte nicht übersehen werden, dass viele der verbundenen Fragestellungen und Herausforderungen nicht grundsätzlich neu sind, sondern vorrangig quantitativ neue Dimensionen erlangt haben bzw. sich in neuen Anwendungskontexten stellen.

Zielsetzung und Vorgehensweise

Ursprüngliche Projektziele

Das Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag (TAB) ist vom Ausschuss für Bildung, Forschung und Technikfolgenabschätzung des Deutschen Bundestages mit dem Projekt »Data-Mining – gesellschaftspolitische und rechtliche Herausforderungen« beauftragt worden. Die ursprüngliche Planung sah vor, anhand von zwei öffentlichen Aufgabenbereichen, für die der Gesetzgeber in besonderem Maße verantwortlich ist, sowohl relevante große Datenbestände als auch die Vielfalt der analytischen Möglichkeiten, die als Data-Mining bezeichnet werden können, anhand von Anwendungsbeispielen darzustellen und die damit einhergehenden Herausforderungen herauszuarbeiten. Zum einen sollten Data-Mining-Ansätze in der Medizin und im Gesundheitssystem vorgestellt werden, die insbesondere die Verwendung personenbezogener Gesundheitsdaten einschließt. Zum anderen sollten Data-Mining-Prozesse unter Verwendung von Geodaten thematisiert werden. Die Annahme war, dass zwei sehr unterschiedliche Anwendungsbereiche erforderlich seien, um gesellschaftlich Herausforderungen in Bezug auf Data-Mining-Prozesse herauszuarbeiten. Folgende leitende Fragestellungen wurden formuliert: Welche traditionellen und welche neuen (öffentlichen) Akteursgruppen können welche Datenbestände aufbauen und für wen sind diese zugänglich? Welche Datenbestände werden bereits heute zu welchen Zwecken zusammengeführt und genutzt? Wem gehören diese Daten, wer entscheidet bezüglich der Verarbeitung und wer kontrolliert die Verfahren? Welche Qualität, welche Verlässlichkeit und Aussagekraft haben Data-Mining-Ergebnisse in analytischer und prognostischer Hinsicht? Wofür und von wem können und sollen Data-Mining-Resultate genutzt werden? Welche neuen Möglichkeiten und Grenzen gibt es bezüglich immer detailgenauerer Erfassungs- und Analyseverfahren? Welche neuen Geschäftsmodelle entstehen?

Im ursprünglichen Projektplan war zudem vorgesehen, neben den Fallstudien ggf. internationale TA- und andere interdisziplinäre Studien zum Thema Big Data/Data-Mining synoptisch auszuwerten, um einen Überblick über die Debatten, Aktivitäten und Einschätzungen zu ethischen Fragestellungen auch in anderen Ländern zu erhalten. Auch eine vertiefte rechtswissenschaftliche Auseinandersetzung zu ausgewählten Aspekten (z. B. zum Eigentums-, Urheber- und Datenschutzrecht) sollte ggf. in Erwägung gezogen werden.

Gutachtenerstellung

Mithilfe von Gutachten sollten zunächst die Datenstrukturen und analysetechnischen Komponenten beim Data-Mining einschließlich der resultierenden Ergebnisse, bestehende nationale Regularien und die damit einhergehenden gesellschaftlichen Herausforderungen anhand von Anwendungsbeispielen untersucht/analysiert und dargestellt werden. Dabei sollte ein Überblick über wesentliche Datenbestände unterschiedlicher Akteursgruppen und ein Ausblick auf sich abzeichnende neue Geschäftsmodelle gegeben werden. Zwei Gutachten wurden zu Projektbeginn in Auftrag gegeben:

- Data-Mining in der Medizin und im Gesundheitssystem − gesellschaftspolitische und rechtliche Herausforderungen. Dr. Josef Schepers, Irene Schlünder, Dr. Johannes Drepper, Sebastian Claudius Semler; TMF – Technologie- und Methodenplattform für die vernetzte medizinische Forschung e. V., Berlin; Dr. Stefan Rüping; Fraunhofer IAIS, Sankt Augustin; Dr. Christoph Quix; Fraunhofer FIT, Sankt Augustin; Dr. Karl Stroetmann, Jonas Rennoch; empirica GmbH, Bonn
- Data-Mining: Gesellschaftspolitische und rechtliche Herausforderungen. Dr. Bodo Bernsdorf, Heide Bierbrauer, Olaf Büscher Andreas Müterthies, Dr. Kian Pakzad, Thomas Wenzel, Sascha Woditsch; EFTAS Fernerkundung Technologietransfer GmbH, Münster

Die Zusammenarbeit mit den Gutachterteams und deren Gutachten schufen die Grundlagen für die weitere Projektarbeit. Den Gutachter/innen sei für ihre Kooperationsbereitschaft bei der mehrmaligen Überarbeitung ihrer Ausführungen und der finalen Erstellung ihrer Gutachten sehr herzlich gedankt.

Im Rahmen dieser aufwendigen Zusammenarbeit wurde u. a. deutlich, dass

- der Begriff Data-Mining je nach fachlicher Perspektive unterschiedlich weit gefasst und unscharf verwendet wird;
- eine Fokussierung auf den bloßen Einsatz datenanalytischer Verfahren zur Mustererkennung (Data-Mining im engeren Sinn) nicht ausreicht, um gesellschaftlich relevante Herausforderungen aufzuzeigen und zu diskutieren;
- infolgedessen eine weitergefasste Prozessbetrachtung (Data-Mining in weiterem Sinn) erforderlich erscheint, wodurch jedoch Aspekte angesprochen

werden, die auch mit der Digitalisierung im Allgemeinen verbunden und nur bedingt Data-Mining-spezifisch sind;
- Data-Mining-Prozesse jeweils einzeln anwendungsbezogen betrachtet werden müssen, um Möglichkeiten und Grenzen der Implementierung derartiger Verfahren in bestehende Strukturen aufzuzeigen.

Die Gutachten zeigten auch, dass ein Überblick über wesentliche Datenbestände unterschiedlicher Akteursgruppen sowie über die Regulierung der Erhebung, Haltung, Bereitstellung, Weiterverwendung und Analyse von Daten in öffentlichen Aufgabenbereichen wegen der geteilten Bund-/Länderzuständigkeit (über)komplex ist. Vor allem im Geodatenbereich ist aufgrund dieser geteilten Zuständigkeit und der zunehmend ubiquitären Erhebung, Bereitstellung und Verwendung von Geodaten ein allgemeiner Überblick über relevante Datenbestände kaum möglich und wenig zielführend. Da der Fokus vorrangig auf öffentlichen Aufgabenbereichen liegt, konnten fundamental neue Geschäftsmodelle kaum herausgearbeitet werden. Herausforderungen ergeben sich vielmehr bei der Datenaufbereitung und -bereitstellung sowie der Integration neuer datenanalytischer Verfahren in die jeweiligen Abläufe öffentlicher Aufgabenbereiche.

Anpassungen im Projektverlauf und Berichtszuschnitt

Die Ausführungen der Gutachten wurden durch umfangreiche eigene Recherchen in mehreren Schleifen überarbeitet und ergänzt. Die hohe Relevanz, die der Digitalisierung im Allgemeinen und datenanalytischen Prozessen im Besonderen derzeit beigemessen wird, führt zu immer neuen Stellungnahmen, Positionierungen, Gutachten und Regulierungsvorschlägen von Kommissionen, Verbänden oder Think-Tanks und zu vielfältigem politischen Engagement – ein sicheres Zeichen, dass neue technologische Möglichkeiten zunehmend anwendungsreif werden und mit gesellschaftlichen Herausforderungen verbunden sind. Die Berücksichtigung der diversen Positionierungen und politischen Aktivitäten erforderte eine kontinuierliche Auseinandersetzung mit dem nach wie vor unscharfen und vielfältig interpretierbaren Oberbegriff Data-Mining. Sie ging zudem mit einem Entscheidungsdilemma einher, einerseits neue Entwicklungen kontinuierlich zu integrieren und andererseits die Arbeiten am Bericht zu beenden. Dies führte zu einer intensiven Befassung mit datenerhebenden und -analysierenden Prozessen im Rahmen des Projekts, aber leider auch zu erheblichen Verzögerungen bei der Berichtsfertigstellung.

Deutlich wurde, dass der Rechtsrahmen für komplexe datenanalytische Prozesse in öffentlichen Aufgabenbereichen vielschichtig und teilweise nur bedingt Data-Mining-spezifisch ist. Die ursprünglich in Betracht gezogene vertiefende Auseinandersetzung mit rechtswissenschaftlichen Fragen sowie ergänzenden Diskursanalysen anderer TA-Einrichtungen hätten die Ausführungen kaum

Data-Mining-spezifisch erweitern können. Die Sondierung vielfältiger TA-Diskurse ergab, dass bei einer Betrachtung auf einer höheren Abstraktionsebene (Datenanalytik im Allgemeinen) ähnliche ethische Aspekte thematisiert werden, u. a. zum Grundrechteschutz, zur Zulässigkeit, zur Transparenz und Kontrollierbarkeit von Datenerhebung und -analyse, zu Verantwortlichkeiten, zu potenziellen Nutzen- und möglichen Risikodimensionen und deren Verteilung. Dazu kann inzwischen auf die Stellungnahme des Deutschen Ethikrats zu Big Data und Gesundheit (Deutscher Ethikrat 2017) und das Gutachten der Datenethikkommission der Bundesregierung (DEK 2019) verwiesen werden.

Data-Mining-spezifische Herausforderungen werden eher in unterschiedlichen Anwendungskontexten sichtbar. In öffentlichen Aufgabenbereichen treffen neue datenanalytische Ansätze auf jeweils national eigenständig regulierte, historisch gewachsene datenverarbeitende Strukturen, die spezifische Besonderheiten mit sich bringen. Statt allgemeine internationale Diskurse zu Big Data, komplexen Algorithmen oder KI synoptisch darzustellen, wird daher im vorliegenden Bericht für einzelne Data-Mining-Anwendungsbeispiele die nationale Vorgehensweise punktuell mit der in ausgewählten anderen Ländern verglichen. Damit soll der Berichtsfokus stärker auf anwendungsbezogene datenanalytische Prozesse sowie die Weiterverwendung/Operationalisierung von Data-Mining-Ergebnissen gelegt werden – zunächst auf einer allgemeineren technischen und rechtlichen Ebene und anschließend spezifischer für medizinische und gesundheitssystemische Kontexte. Denn in diesen Bereichen werden komplexen Datenanalysen nahezu unisono besondere Potenziale unterstellt. Zugleich attestieren unterschiedliche vergleichende Studien dem nationalen Gesundheitssystemeinen erheblichen Entwicklungsbedarf bei der Digitalisierung vielfältiger Prozesse (z. B. BMWi 2018; svr Gesundheit 2021; Thiel et al. 2018). Ein Einblick in die bestehenden Strukturen und Aufgabenspektren beteiligter Akteursgruppen soll die Auseinandersetzung mit den Chancen und Herausforderungen komplexer datenanalytischer Prozesse fundieren. Zudem können ethische Aspekte zum Grundrechteschutz betroffener Personen, zur Verantwortlichkeit, Nachvollziehbarkeit und Kontrollierbarkeit datenanalytischer Prozesse sowie der Nutzenbewertung daraus hervorgehender Ergebnisse anwendungsbezogen thematisiert werden, denn diese sind in der Medizin und im Gesundheitssystem seit langem in besonderem Maße normativ verankert.

Auf eine separate Fallstudie zu den Möglichkeiten, Grenzen und Herausforderungen komplexer raumbezogener Datenanalysen wird in diesem Bericht verzichtet. Das hat mehrere Gründe: Die Ausführungen des Gutachtens von Bernsdorf et al. (2015) zu bereits weitgehend normierten Geodaten, zu den analytischen Verfahren der datenbasierten räumlichen Mustersuche sowie zu den gesetzlichen Vorgaben der Bereitstellung von Geodaten aus öffentlichen Aufgabenbereichen sind ein wesentlicher Bestandteil der technischen und rechtlichen Grundlagenkapitel dieses Berichts. Bezüglich des Ausbaus nationaler

Geodateninfrastrukturen ist Deutschland im europäischen Vergleich sehr gut positioniert (Cetl et al. 2017). Die Aktualisierung der Geo(basis)daten und die Georeferenzierung vielfältiger Fachdaten auf kommunaler, Landes- und Bundesebene und deren Bereitstellung über diese Infrastrukturen ist eine Daueraufgabe. Dem Bundestag wird regelmäßig darüber berichtet.[3] Geodaten sind zunehmend ubiquitär verfügbar und raumbezogene Analysen sind eine Querschnittsaktivität in vielfältigen Bereichen. Einige Data-Mining-Anwendungsbeispiele werden sowohl in den Grundlagenkapiteln als auch in der gesundheitssystemischen Fallstudie diskutiert. Zudem wurden Möglichkeiten und Grenzen, Chancen und Risiken sowie allgemeine gesellschaftliche Herausforderungen datenbasierter räumlicher Struktur- und Mustersuchen auch schon in diversen anderen TAB-Arbeitsberichten ausführlich anwendungsbezogen thematisiert.[4] Eine weitere Auseinandersetzung mit derartigen Verfahren im Rahmen dieser Überblickstudie hätte nach Einschätzung des TAB kaum substanziellen Mehrwert geboten.

Der vorliegende Endbericht bringt die Ausführungen der Gutachten, die eigenen Recherchen und die daraus gewonnenen Schlussfolgerungen zusammen. Die Verantwortung für die Auswahl, Strukturierung und Verdichtung des Materials liegt bei der Verfasserin dieses Berichts, Dr. Katrin Gerlinger. Dank geht an Dr. Alma Kollek, Dr. Christoph Revermann und Dr. Arnold Sauter, die durch Gegenlesen und detailliertes Kommentieren zur Verbesserung des vorliegenden Berichts entscheidend beigetragen haben sowie an Carmen Dienhardt und Brigitta-Ulrike Goelsdorf für die sorgfältige Durchsicht des Manuskripts, die Bearbeitung der Abbildungen und die Erstellung des Endlayouts.

Berichtsaufbau

Der Begriff Data-Mining wird in diesem Bericht weit ausgelegt und umfasst nicht nur die unmittelbare Anwendung mathematisch-statistischer Verfahren auf Datensätze oder -bestände (Data-Mining im engeren Sinn), sondern den gesamten Prozess der datenbasierten Mustererkennung und Regelableitung bzw. Informations-/Erkenntnisgewinnung (Data-Mining im weiteren Sinn). Dessen

3 Geo-Fortschrittsberichte (Bundesregierung 2005, 2008, 2012a, 2017 u. 2021b)

4 Fernerkundung: Anwendungspotenziale in Afrika (TAB 2012); Digitalisierung der Landwirtschaft (TAB 2021); Beobachtungstechnologien im Bereich der zivilen Sicherheit – Möglichkeiten und Herausforderungen (TAB 2022a); Innovative Technologien, Prozesse und Produkte in der Bauwirtschaft (TAB 2022b); Künstliche Intelligenz und Distributed-Ledger-Technologie in der öffentlichen Verwaltung (TAB 2022c); Chancen und Risiken der Digitalisierung kritischer kommunaler Infrastrukturen an den Beispielen der Wasser- und Abfallwirtschaft (www.tab-beim-bundestag.de/projekte_chancen-und-risiken-der-digitalisierung-kritischer-kommunaler-infrastrukturen-an-den-beispielen-der-wasser-und-abfallwirtschaft.php; 10.01.2022).

besonderes Potenzial besteht darin, dass ausreichend valide Ergebnisse und analytische Vorgehensweisen generalisiert und in neuen Situationen des gleichen Sachverhalts angewendet werden können. Die (analyse)technischen Komponenten und Vorgehensweisen sowie vielfältige normative Aspekte bei Data-Mining-Aktivitäten werden im Bericht stufenweise erschlossen.

In Kapitel 2 werden zum einen die für Data-Mining notwendigen maschinenlesbaren Daten, deren Bestandshaltung und Bereitstellung und zum anderen die einzelnen Schritte des Data-Mining-Prozesses bis zur Entwicklung algorithmischer Systeme aus (informations)technischer Perspektive dargestellt. Ein historisches Beispiel soll den Einstieg in die Thematik veranschaulichen und zeigen, dass die datenbasierte Mustererkennung und die Informationsableitung keineswegs fundamental neu sind, auch wenn das, was einzelne Menschen früher manuell bzw. intellektuell erledigten, inzwischen in viel größerem Umfang durch die Verknüpfung unterschiedlicher Datenbestände mittels diverser Analysetechniken und leistungsstarker Computerarchitekturen weitgehend maschinell realisiert werden kann.

In Kapitel 3 werden rechtliche Aspekte zum Umgang mit Daten und den aus Data-Mining-Prozessen resultierenden Ergebnissen umrissen. Dazu gehören die schutzwürdigen Interessen sowohl von betroffenen Personen bei personenbeziehbaren Daten (durch die DSGVO definiert) als auch von datenerhebenden Stellen (teilweise durch das Urheberrecht definiert) sowie die sich daraus ergebenden Data-Mining-Möglichkeiten und -Grenzen. Die DSGVO und das Urheberrecht (Anhang 1) ermöglichen Datenweiterverwendungen zu Forschungszwecken und Data-Mining. Deren Resultate können zunehmend weiterentwickelt und gewinnbringend vermarktet werden, ohne dass sie spezifischer Regulierung unterworfen sind. Vor allem bei risikoreichen algorithmischen Entscheidungs(unterstützungs)systemen wird eine stärkere Regulierung zunehmend gefordert. Diesbezüglich bietet das Medizinprodukterecht möglicherweise Regulierungsoptionen. Damit wird die Brücke zur vertiefenden Fallstudie »Data-Mining in der Medizin« geschlagen.

In Kapitel 4 werden zunächst die rechtlichen und technischen Besonderheiten der Erhebung, Haltung und Analyse medizinischer Daten dargestellt. Diese Daten sind die Basis für die Entwicklung von Scoringverfahren, prädiktiven Modellen und Bilderkennungsverfahren, die in einem weiteren Schritt zu algorithmischen Entscheidungs(unterstützungs)systemen ausgebaut werden können. Anhand unterschiedlicher Beispiele wird gezeigt, welche Herausforderungen die in der Medizin notwendigen Sicherheits-, Leistungs- und Nutzennachweise mit sich bringen und wie aufwendig der Weg derartiger Ergebnisse aus Data-Mining-Prozessen in die Regelversorgung ist. Medizinische Einrichtungen müssen zudem für vielfältige administrative Aufgaben regelmäßig standardisierte Datensätze zusammenstellen und an unterschiedliche Institutionen der gesundheitssystemischen Selbstverwaltung übermitteln.

In Kapitel 5 werden einige dieser Institutionen mit ihren jeweiligen datenbezogenen Aufgaben und den bei ihnen entstehenden Datenbeständen vorgestellt. Auch diesen Versorgungsdaten werden große Data-Mining-Potenziale unterstellt. Anhand weiterer Anwendungsbeispiele werden Möglichkeiten und Grenzen diskutiert, aus diesen Daten Informationen zu gesundheitssystemischen Herausforderungen zu extrahieren.

Das abschließende Kapitel 6 fasst wesentliche Punkte zum nach wie vor vielfältig interpretierbaren Data-Mining-Begriff zusammen. Die Fallstudien zu Data-Mining-Prozessen in der Medizin und im öffentlichen Gesundheitssystem zeigen, dass vor allem die Datenbereitstellung und der Umgang mit den aus Data-Mining resultierenden Informationen und Algorithmen gesellschaftliche Herausforderungen mit sich bringen.

2 Data-Mining aus analytisch-technischer Sicht

2.1 Data-Mining – was ist das?

Der Begriff Data-Mining wird seit Anfang der 1990er Jahre verwendet (z. B. Fayyad et al. 1996; Frawley et al. 1992), zunächst vorwiegend aus analytischer, dann auch aus softwaretechnischer und anwendungsorientierter Sicht (umfassender verfahrenstechnischer Einblick z. B. in Witten et al. 2011). Aus analytischer Sicht wird Data-Mining mit der Anwendung von unterschiedlichen mathematisch-statistischen Verfahren assoziiert, um in Datenbeständen neue, potenziell nützliche Strukturen und Muster zu identifizieren (Schepers et al. 2015, S. 32). In diesem Verständnis wird Data-Mining assoziiert mit einem Prozess, den Fayyad et al. (1996) als »Knowledge Discovery in Databases« bezeichneten. Eine wie auch immer geartete größere Datenbasis ist eine notwendige Bedingung, Daten einzelner oder weniger Datenobjekte reichen dazu in der Regel nicht aus. Dieses datenbasierte Knowledge Discovery hat erhebliche Schnittmengen mit den neueren Schlagworten Big oder Smart Data, maschinelles Lernen oder KI.

Im Kern zielen alle mit diesen Schlagworten assoziierten Vorgehensweisen darauf ab, in großen, nur noch maschinell verwalt- und verarbeitbaren Datenmengen Strukturen und Muster zu erkennen, Informationen abzuleiten und durch Regeln zu generalisieren, die auf neue Situationen zu übertragen und angewendet werden können. Eine Abgrenzung der analytischen Verfahren, die eher zu dem einen oder zu einem anderen Schlagwort gehören, ist schwierig. Rückblickend scheint der Begriff Data-Mining eher in der mathematischen Statistik generiert worden zu sein, wohingegen die Begriffe maschinelles Lernen und künstliche Intelligenz eher in der Informatik entstanden sind (DEK 2019, S. 59; Witten et al. 2011, S. 28 f.). Die Begriffe Big und Smart Data betonen stärker die kontinuierlich wachsenden Datenmengen, die im Zuge der Digitalisierung zahlreicher Alltags- und Geschäftsprozesse entstehen, sowie die darin enthaltenen Informationen, die nur noch maschinell extrahiert werden können. Witten et al. (2011) plädieren dafür, keine Trennlinie zwischen diesen Begriffen zu suchen, sondern sie eher als eine Art Kontinuum aufzufassen, zumal gleiche Prozessabläufe entstanden und oft ähnliche Verfahren und Algorithmen eingesetzt werden. Auch bei Betrachtungen zu gesellschaftlichen Chancen und Herausforderungen ergeben sich bei all diesen unscharfen Begriffen viele Gemeinsamkeiten. Aus dieser Perspektive wird mitunter empfohlen, statt von Data-Mining besser von *komplexen Datenanalysen* zu sprechen (Schepers et al. 2015, S. 20 ff.; Triaille et al. 2014, S. 9 f.). Dieser nüchterne Begriff biete zudem eine

Möglichkeit, die philosophischen Auseinandersetzungen zu den Begriffen Lernen bzw. maschinelles Lernen, Formen der Wissenserweiterung oder der (künstlichen) Intelligenz sowie zwischen menschlichen Fähigkeiten und technischen Möglichkeiten zu umgehen.[5]

Weitgehend übereinstimmend werden mit Data-Mining mathematisch-statistische Verfahren verbunden, die in Datenbeständen strukturelle Muster (Ähnlichkeiten, Zusammenhänge, Unterschiede) erkennen und darstellen (anhand von Parametern, Formeln, Entscheidungsregeln, mathematisch-statistischen Modellen). Diese Muster und deren Darstellungen können in wissenschaftlichen Auseinandersetzungen diskutiert und hinterfragt und sofern sie verallgemeinerbar sind, auf neue Situationen übertragen und angewendet werden. Text- oder Web-Mining sowie Bilderkennung gelten als Spezialbereiche für besondere Datentypen und -quellen (TAB 2014, S. 43; Witten et al. 2011, S. 3 ff.). Teilweise werden mit Data-Mining nur sekundäre Analysen von Daten, die in anderen Kontexten entstanden sind (z. B. Tracking des Internetverhaltens durch Analysen von Verkehrsdaten), assoziiert oder nur ausgewählte strukturerkennende Verfahren darunter gefasst (Knobloch/Weidner 2000, S. 346). Im Rahmen dieser Untersuchung werden komplexe und aufwendige datenanalytische 0,,,,,,,Verfahren zur Erkennung von Strukturen und Mustern in Datenbeständen und zur Ableitung von Informationen mit dem Begriff Data-Mining assoziiert und einerseits triviale statistische Verfahren und andererseits vollautomatisierte Prozesssteuerungen davon abgegrenzt. Diese Grenzziehung ist jedoch dynamisch: Was vor Jahren als komplex und aufwendig galt, kann durch den technischen Fortschritt trivial einfach werden, und Softwareprogramme, die zunächst lediglich einzelne Berechnungen durchführten, können im Laufe der Entwicklung immer umfangreichere und komplexere Aufgaben automatisiert bewältigen (Bernsdorf et al. 2015, S. 36; Schepers et al. 2015, S. 20 ff.). Insbesondere bei der Analyse großer Datenmengen erscheinen eingesetzte Verfahren mitunter nur deshalb komplex, weil auf unterschiedliche Datenbestände zugegriffen wird und Analyseschritte zerlegt und umfänglich parallel ausgeführt werden. Die eigentlichen mathematisch-statistischen Verfahren gibt es jedoch teilweise seit langem, z. B. zur Klassifikation, Zusammenfassung, Gruppierung oder Ausreißerkennung.

Insbesondere wenn der Frage nach den mit den Analysetechniken verbundenen gesellschaftlichen Chancen und Herausforderungen nachgegangen werden soll, liegt es auf der Hand, nicht nur den unmittelbaren Einsatz von mathematisch-statistischen Verfahren zur Erkennung von (neuen) Strukturen und Mustern in Datenbeständen (*Data-Mining im engeren Sinn*), sondern den gesamten Prozess des Knowledge Discovery in Databases zu betrachten (*Data-*

5 Diese philosophische Debatte wurde im TAB-Arbeitsbericht »Technologien und Visionen der Mensch-Maschine-Entgrenzung« aufgegriffen (TAB 2016b). Sie wird in diesem Bericht nicht vertieft.

Mining im weiteren Sinn) (Bernsdorf et al. 2015, S. 36; Schepers et al. 2015, S. 31). Im Rahmen dieser Untersuchung werden folgende Prozessschritte mit Data-Mining im weiteren Sinn assoziiert:[6]

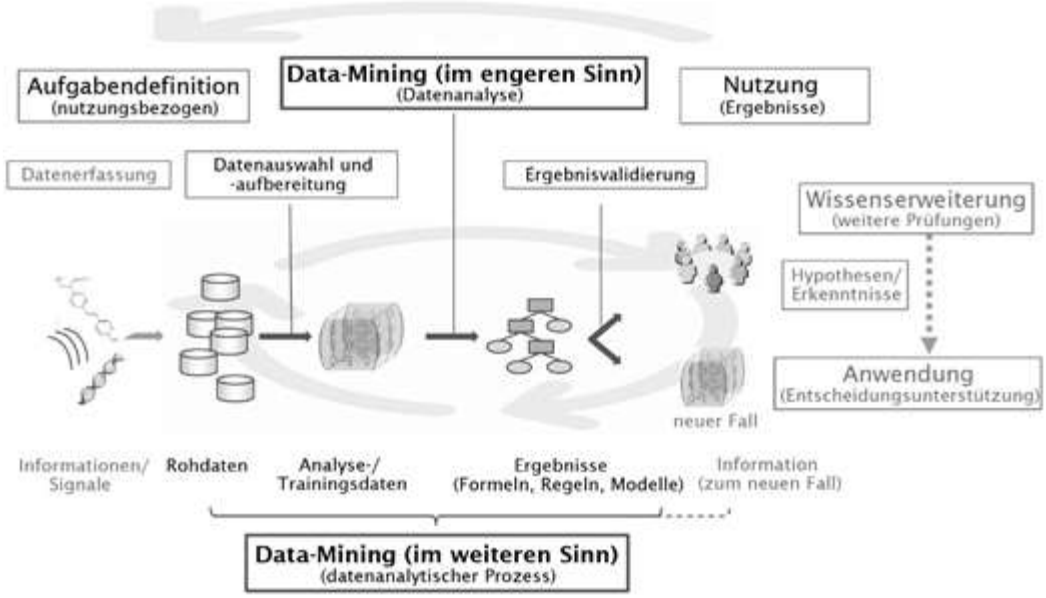
- *Aufgabendefinition bzw. Spezifikation des Untersuchungsauftrags*: Data-Mining ist eine im weiteren Sinn zweck- oder nutzungsgetriebene Analyse von Daten. Sie beginnt mit der Formulierung eines Untersuchungsziels, das sich aus einem anwendungsbezogenen Kontext ergibt (z. B. Suchen von Gemeinsamkeiten oder Auffälligkeiten, Klassifizierung oder Gruppierung von Objekten, Ableitung von Prognosen).
- *Datenauswahl und -aufbereitung*: Je nach Aufgabe sind aus oftmals unterschiedlichen Datenbeständen geeignete Teile auszuwählen, mitunter ist die Nutzungsberechtigung zu prüfen), die ausgewählten Daten werden bereinigt (z. B. ist der Umgang mit fehlenden und fehlerhaften Werten zu klären) und aufbereitet (z. B. Merkmale/Variablen umrechnen, transformieren oder zusammenfassen).
- *Datenanalyse*: Je nach Aufgabe und Datentypen (Zahlen, Orts-/Zeitangaben, Zeichenketten/Texte, Bilder) kommen unterschiedliche mathematisch-statistische Verfahren in Betracht, die mittels Algorithmen und Software auf dem aufbereiteten Analysedatensatz ausgeführt werden und Ergebnisse liefern.
- *Ergebnisvalidierung*: Analyseergebnisse werden auf unterschiedliche Weise verfahrensintern und/oder verfahrensextern geprüft und bewertet.

Diese Data-Mining-Prozessschritte sind in Abbildung 2.1 grafisch dargestellt. Auch der Prozess des Data-Mining im weiteren Sinn baut zum einen auf unterschiedliche Verfahren der vorgelagerten Datenerfassung und -speicherung auf. Zum anderen können deren Ergebnisse auf unterschiedliche Art und Weise genutzt werden. Sie werden in der Regel fachlich inhaltlich diskutiert, um deren allgemeine Gültigkeit bzw. Generalisierbarkeit zu untermauern bzw. eine spezifische Bewertung und Validierung ermittelter Ergebnisse zu ermöglichen sowie Hypothesen abzuleiten, Erkenntnisse zu fundieren oder in Frage zu stellen und bestehendes Wissen zu erweitern (wissenschaftliche Verwendung). Ausreichend valide Regeln und Modelle können auch auf neue Situationen übertragen und angewendet werden, um je nach Aufgabendefinition, diese zu klassifizieren, Prognosen zu erstellen und dadurch situativ neue Informationen bzw. Daten zu generieren (operative Anwendung). Operationalisierbare Regeln, Modelle

6 Fayyad et al. (1996, S. 40) schlüsselte den Data-Mining-Prozess aus verfahrenstechnischer Sicht etwas differenzierter auf. Die für diesen Bericht definierten Prozessschritte lehnen sich stärker an das um die Jahrtausendwende von Shearer (2000) entwickelte und unter Datenanalytikern bekanntere CRISP-DM-Modell (Cross Industry Standard Process of Data Mining) an (ausführlicher z. B. in Schepers et al. 2015, S. 31 ff.). Bei der Befassung mit den gesellschaftlichen Chancen und Herausforderungen von Data-Mining ist eine differenzierte Prozessbetrachtung aus Sicht des TAB nicht erforderlich.

und Algorithmen können als neue Funktionalitäten in bestehende Software und Services integriert oder als eigenständige Informations-, Assistenz- oder Entscheidungsunterstützungssysteme genutzt und verwertet werden. Es gibt unterschiedliche Meinungen, inwiefern diese vor- und nachgelagerten Prozesse ebenfalls zum Data-Mining-Prozess gehören. Folgen derartigen Vorgehens und gesellschaftliche Herausforderungen lassen sich jedoch nur herausarbeiten und abschätzen, wenn der Gesamtprozess in den Blick genommen wird.

Abb. 2.1 Data-Mining: schematische Darstellung der Prozessschritte



Eigene Darstellung

Wenn Data-Mining als automatisierte Analyse definiert wird, bezieht sich diese Definition auf Data-Mining im engeren Sinn. Betrachtet man den gesamten Prozess (Data-Mining im weiteren Sinn) wird deutlich, dass ein hoher Automatisierungsgrad bisher vor allem bei der eigentlichen Datenanalyse möglich ist, weil die einzelnen Schritte spezifischer mathematisch-statistischer Verfahren durch Algorithmen und Software ausgeführt werden. Bei der Aufgabendefinition, der Datenauswahl und -aufbereitung wie auch bei der Ergebnisprüfung und -validierung sowie der Informations- und Wissensableitung sind nach wie vor viele gedankliche und manuelle Arbeitsschritte erforderlich (Schepers et al. 2015, S. 33 f.). Auch der Übergang von einem Prozessschritt zum nächsten ist bisher nicht automatisiert. Vielfach werden (Zwischen-)Ergebnisse geprüft, Eingangsparameter angepasst, Daten nach und nach hinzugezogen oder ausgeschlossen und einzelne Schritte wiederholt, bis Resultate als richtig, valide oder nützlich eingestuft werden können. Wenn dieser Punkt erreicht ist, kann eine

operative Anwendung ermittelter Regeln in neuen Situationen in Erwägung gezogen werden. Ein historisches Beispiel soll zunächst den Data-Mining-Prozess und die Weiterverwendung der Ergebnisse veranschaulichen.

Historisches Beispiel: Choleraepidemie in London 1854

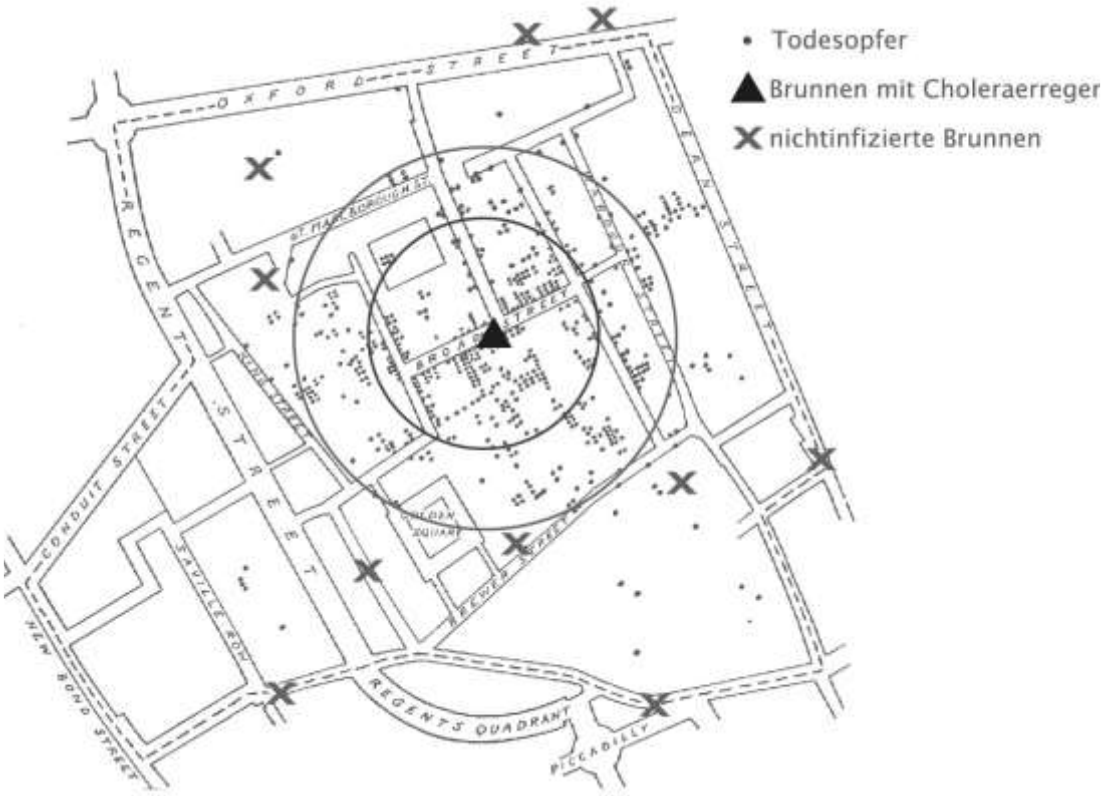
Als historisches Beispiel für Data-Mining wird oftmals eine Datenanalyse des britischen Arztes John Snow aus dem 19. Jahrhundert herangezogen (Bernsdorf et al. 2015, S. 47; Schepers et al. 2015, S. 27): Er bezweifelte, dass die Ursache der Cholera mit der bis dato unter Ärzten gängigen Miasmentheorie (Infektionskrankheiten würden durch üble Dünste oder in der Luft zirkulierende faulige Stoffe übertragen, die aus dem Boden entweichen oder aus Gewässern kommen könnten) erklärt werden kann. Snow vermutete, dass das Übertragungsmedium von Cholera verunreinigtes Trinkwasser sei. Als im Londoner Stadtteil Soho 1854 eine Choleraepidemie ausbrach, sammelte er vielfältige Informationen zu den Choleraopfern, u. a. zu deren Wohnorten, die er auf einer Karte des Stadtteils markierte (Punkte in Abb. 2.2). Zudem markierte er in dieser Karte auch die Positionen der örtlichen Wasserbrunnen (B1 bis B9 in Abb. 2.2).

Allein durch die grafische Darstellung wurde die Häufung der Todesopfer um ein Zentrum offensichtlich (Mustererkennung). In diesem Zentrum stand der Broad-Street-Brunnen (B5). Snow interpretierte sein Analyseergebnis nicht nur als Beziehung zweier Sachverhalte (Wasserbrunnen und Choleraopfer), sondern unterstellte eine Ursache-Wirkungs-Beziehung. Obwohl auch in weiterer Entfernung vom Brunnen einige Todesfälle zu verzeichnen waren und Snow zu diesem Zeitpunkt keine biologisch dezidierte Begründung liefern konnte, forderte er die sofortige Stilllegung des Brunnens (Ableitung einer operativen Entscheidung). Wahrscheinlich stieß er nicht auf allseits offene Ohren, widersprach er doch den vorherrschenden Meinungen zur Krankheitsübertragung. Überliefert ist, dass Snow den Pumpenschwengel in der Broad Street eigenmächtig abmontierte und so die Nutzung des Brunnens unterband, woraufhin die Zahl der Choleraopfer sank (Gerste 2014, 2020).

Was hat Snow analytisch getan? Er hat zunächst eine Beziehung (Korrelation) zwischen zwei Sachverhalten/Datenobjekten (Choleraopfer und Wasserbrunnen) über ein verbindendes Merkmal (Standort) anhand der jeweiligen Merkmalsausprägungen (jeweilige Adresse der Choleraopfer und Position der Wasserbrunnen) hergestellt. Was er 1854 mit Zettel und Stift manuell vollzog, bezeichnet man heute als räumliches Clustern und Hotspot-Analyse. Die zur Interpretation von Geodaten nach wie vor wichtige Visualisierung ist eine Überlagerung von Geobasisdaten (Londoner Stadtplan) mit spezifischen Geofachdaten (Standorte der Brunnen und Wohnorte der Choleraopfer). Derartiges Vor-

gehen wird heute mittels spezieller Software realisiert, die auf große Datenbestände und standardisierten Algorithmen zugreift (auch als Geoinformationssysteme bezeichnet).

Abb. 2.2 Position der Choleraopfer und der Wasserbrunnen in London 1854



Quelle: Bernsdorf et al. 2015, S. 47 nach https://de.wikipedia.org/wiki/John_Snow_(Mediziner) (8.12.2021)

Um seine von der damaligen Mehrheitsmeinung abweichende Hypothese zur Ursache-Wirkungs-Relation und seine eigenmächtige operative Handlung zu fundieren, ging Snow methodisch einige Schritte weiter. Zum einen hinterfragte vermeintliche Datenausreißer und stellte fest, dass vom Brunnen weiter entfernte Opfer Wasser aus diesem Brunnen getrunken und Häuser in Brunnennähe ohne Choleraopfer eine eigene Wasserversorgung hatten. Damit konnte er die Kausalität von Brunnen und Choleraerkrankung erhärten. Zum anderen nahm er Stuhl- und Wasserproben. Mit seinen damaligen Möglichkeiten konnte er die Choleraerreger als eigentliche biologische Ursache der Cholera nicht zweifelsfrei nachweisen. Diese biologischen Beweise lieferte Robert Koch 30 Jahre später. Rückblickend fällt Snows Vorgehen in die Zeit, in der der medizinische Fokus auf überindividuelle Ebenen erweitert wurde (Public-Health-Ansatz) und

die datengestützte Forschung an Bedeutung gewann. Er selbst gilt als einer der Pioniere räumlicher Datenanalysen und Mitbegründer der Epidemiologie.

Was ist heute anders, was nicht?

Die heute maschinell generierten und gespeicherten Datenbestände haben mehrheitlich eine viel höhere Detailgenauigkeit (Granularität) als früher und fallen zugleich in zunehmend großer Menge an, sind in der Summe extrem heterogen und nicht immer fehlerfrei. Für einzelne Fragestellungen sind oft nur einzelne Segmente großer Datenbestände erforderlich, teilweise können dafür auch nur bedingt spezifische Daten genutzt werden. Wichtig ist zumeist, dass sie schnell verarbeitet werden und sofort Ergebnisse liefern können (Big-Data-Konzept z. B. in Holzinger/Jurisica 2014, S. 4; Wiegerling et al. 2018, S. 2). Solche Datenmengen sind für den Menschen selbst kaum noch unmittelbar erfassbar und verständlich. Es gibt jedoch kontinuierlich leistungsstärkere und nutzerfreundlichere Informationstechnologien, mit denen Daten immer einfacher erfasst, dauerhaft gespeichert, bereitgestellt und analysiert werden können. Zudem hat der Schutz der Daten und der Rechte beteiligter Personen heute einen viel höheren Stellenwert.

Datenanalytiker/innen weisen darauf hin, dass auch mit immer nutzerfreundlicheren Informationssystemen und zunehmender Automatisierung einzelner Prozessschritte für Data-Mining-Aktivitäten nach wie vor erhebliche Fachkenntnisse erforderlich sind, um spezifische Analysefragen als mathematisch-statistische Probleme zu definieren, die dafür erforderlichen Daten und Verfahren problemorientiert auszuwählen, interne Gütekriterien ggf. einzuordnen, Ergebnisse zu prüfen und zu bewerten, Grenzen der Analysen und Fehler zu erkennen sowie falsche Schlüsse zu vermeiden (Knobloch/Weidner 2000, S. 354). Nichtdurchdachte Data-Mining-Untersuchungen können bedeutungslose, irreführende oder falsche Strukturen und Muster hervorbringen – mitunter auch als Data-Dredging bezeichnet (in freier Übersetzung auch als Schlammbaggern bezeichnet) (Bernsdorf et al. 2015, S. 37).

Einen Unterschied gibt es mitunter bei den mit neuen wahrscheinlichkeitstheoretischen analytischen Verfahren ermittelten Ergebnissen und deren Darstellung. Im Unterschied zu klassischen mathematisch-statistischen Verfahren (z. B. Cluster- oder Regressionsanalysen) werden die aus Trainingsdaten ermittelten Beziehungen zwischen Sachverhalten, resultierende Entscheidungsstrukturen und Prognosemodelle bei neuronalen Netzen nicht mittels Kennziffern, Regeln oder Formeln ausgewiesen (Kap. 2.3.2). Ob ein solches Vorgehen, das keine Entscheidungsstrukturen offenlegt, noch Data-Mining im Sinne des Knowledge Discovery in Databases ist, oder nur künstliche Intelligenz, weil lediglich Algorithmen trainiert werden, definierte Aufgaben zu lösen, diese Algorithmen dann in neuen Situationen (Fach-)Menschen Informationen anbieten,

ohne dass letztere ihr datenanalytisches Wissen selbst erweitern, scheint eher eine philosophische Frage zu sein. Um jenseits dieser Debatte (ausführlicher z. B. in TAB 2016) die gesellschaftspolitischen und rechtlichen Herausforderungen komplexer Datenanalysen schrittweise zu erschließen, werden nachfolgend sowohl der Umgang mit Daten als auch die Analysetechniken und die Nutzung resultierender Ergebnisse ins Zentrum der Betrachtung gestellt.

2.2 Daten: Formen, Strukturen und Bereitstellung

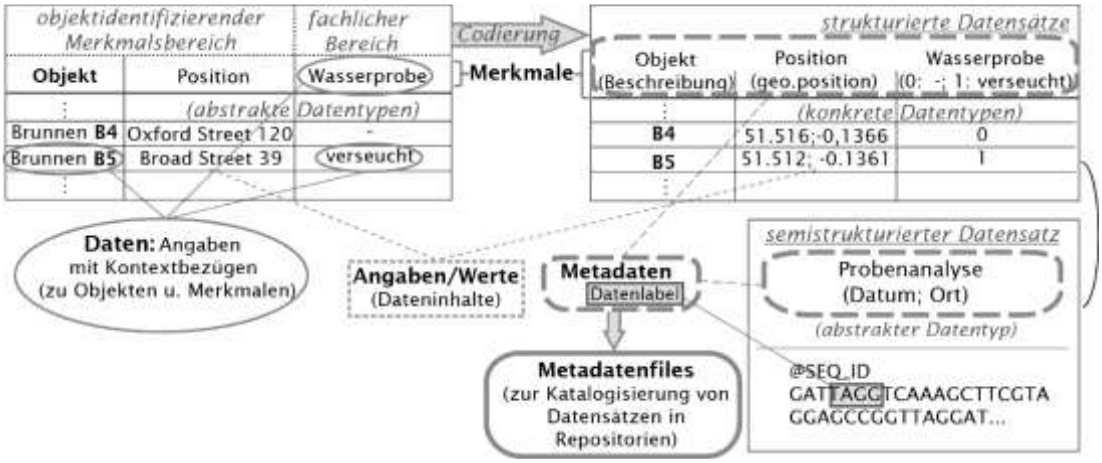
Obwohl Daten heute nahezu allgegenwärtig sind, gibt es bisher keine allgemeingültige Datendefinition. Teilweise werden die Begriffe Daten und Informationen synonym verwendet, teilweise sind Daten eine Oberkategorie für diverse Aufzeichnungen (DEK 2019, S. 52). Ohne den Anspruch zu erheben, die bestehende kategoriale Unbestimmtheit grundsätzlich aufzulösen zu können, soll der Datenbegriff für den nachfolgenden Bericht zunächst aus informationstechnischer und im Anschluss in Kapitel 3 aus rechtlicher Perspektive konkretisiert werden.

2.2.1 Wesensmerkmale und Formen

Aus informationstechnischer Sicht sind Daten (maschinen)lesbare Informationen. Um sie maschinell verarbeiten und analysieren zu können, müssen sie eine gewisse Minimalstruktur erhalten. Dazu werden die inzwischen oft mit technischen Hilfsmitteln (Sensoren, Messgeräte, Kameras) primär erhoben Messwerte als *Angaben/Werte* (Dateninhalte) deklariert und mit Kontexten verbunden, d. h., sie erhalten Referenzen zu *Objekten* und werden durch *Merkmale* strukturiert dargestellt (Abb. 2.3 links).

Dateninhalte, Objekte und Merkmale können vielfältig klassifiziert, gruppiert und strukturiert werden. In diesem Kontext sind Features oder Klassen abstrahierte Fachobjekte (z. B. können Choleraopfer eine Klasse unter den Todesfällen einer Region bilden). Attribute sind Eigenschaften von Objekten oder Features/Klassen, die diesen über ein Merkmal zugeordnet werden (Bernsdorf et al. 2015, S. 30). Im Beispiel wäre u. a. die Wasserprobe (Merkmal) des Brunnens B5 (Objekt) der Londoner Wasserversorgung (Feature) verseucht (Attribut) mit Vibrio cholerae (Angabe). Bei komplexen semistrukturieren Datensätzen (z. B. genetischen Daten, Bilder, Videos) bezeichnet man solche Strukturierungen auch als *Datenlabeling*. Dazu gehören u. a. Markierungen auf Bildsegmenten (Annotationen, Tags), Schlüsselwörter bei Texten, Kategorisierungen von Videos nach ihrem Inhalt. Dies ist ein wichtiges Element der Datenaufbereitung für komplexe Datenanalysen (Training künstlicher neuronaler Netze mit maschinellen Lernverfahren; Kap. 2.3.2).

Abb. 2.3 Datenstrukturen



Eigene Darstellung

Derartige Zusatzangaben zu Objekten/Features, Merkmalen/Attributen und Klassifikationen werden auch als *Metadaten* (Daten über Daten) bezeichnet. Sie können vielfältige technische, administrative oder organisatorische Kontextangaben liefern z. B. zum Messverfahren, zur Bezeichnung, Darstellung und Codierung, mitunter zu Zeit, Ort und/oder Objekten. Durch die Digitalisierung werden vielfältige Prozessschritte automatisiert in Logfiles immer detaillierter dokumentiert, wodurch immer umfangreichere Metadatenbestände entstehen. Teilweise werden diese Metadaten auch als Rand- oder Verkehrsdaten bezeichnet. Etliche (Zusatz-)Angaben wie z. B. Zeit- oder Ortsangaben können in einer Datenstruktur als Merkmal direkter Bestandteil des Datensatzes sein und in einer anderen Darstellung einen Datensatz ergänzend beschreiben. Auch Metadaten müssen für eine maschinelle Datenverarbeitung in strukturierter Form vorliegen. Für Data-Mining-Aktivitäten sind Metadaten in mehrfacher Hinsicht relevant:

- Über die Metadatenfiles werden Datensätze in umfangreichen Repositorien katalogisiert (ähnlich der Registerkarte einer Publikation in einem Bibliothekskatalog; ausführlicher z. B. EK 2013). Über die Kataloge können Datensätze maschinell gesucht, visualisiert und ggf. extrahiert werden.
- Für die Programmierung der Ausführungsskripte von Datenanalysen und/oder die Parametrisierung von Algorithmen reicht es zunächst zu wissen, wie Merkmale/Variablen bezeichnet und Dateninhalte codiert sind, ohne dass ein Zugang zu den Dateninhalten erforderlich ist.
- Metadaten können selbst Gegenstand von Datenanalysen sein (insbesondere Verkehrsdaten, die Segmente des Verhaltens von Personen aufzeichnen und die Erstellung von individuellen Verhaltensprofilen, anhand derer Personen identifiziert werden könnten, technisch ermöglichen; Kasten 3.7).

Aufgrund der hohen datenanalytischen Relevanz von Metadaten zeichnet Pasquinelli (2018) das Bild einer »Gesellschaft der Metadaten«, die mittels Data-Mining auch neue Formen der Überwachung oder der Verhaltenssteuerung hervorbringen kann. Zunehmend wird darauf hingewiesen, dass situativ untersucht werden müsse, welche gesellschaftlichen Folgen mit dem jeweiligen datenanalytischen Ansatz einhergehen.

Im Rahmen dieser Untersuchung werden Daten als Angaben/Informationen zu Objekten (z. B. Situationen, Gegenständen, Ereignissen aber auch Personen als rechtlich besondere Datenobjekte; Kap. 3.1) aufgefasst, die maschinenlesbar zumindest temporär gespeichert und verarbeitet werden. Sie haben folgende wesentliche Eigenschaften:

- Dateninhalte werden zwar kontinuierlich detailgenauer (z. B. können genetische Daten bis auf molekularer Ebene erhoben werden), dennoch bilden sie Objekte und Realitäten nicht in voller Komplexität ab.
- Daten sind von unterschiedlicher Qualität und nicht immer fehlerfrei.
- Daten werden strukturiert gespeichert. Die Strukturierung hat eine semantische (betrifft die Dateninhalte und die dafür verwendeten Terminologien und Codierungen; Kasten 2.1) und eine syntaktische Ebene (betrifft die Anordnung, Darstellung und Formate u. a. Tabellen, Texte, Bilder, Videos, Genomsequenzen). Dadurch werden sie maschinell verarbeitbar.
- Daten können vervielfältigt und gelöscht, über die Kontextbezüge und die Strukturierung verknüpft, erweitert und verändert werden (u. a. können Dateninhalte sowie Datenobjekte und Merkmale klassifiziert, gruppiert und zusammengefasst werden).
- Aufgrund der Maschinenlesbarkeit können Daten informationstechnisch vielfältig verknüpft und analysiert werden.

In diesem Sinne sind Daten(sätze) mehr als bloße Ziffernfolgen – das *Mehr* liegt in den jeweiligen Strukturen und den Bezügen. Diese Strukturen mit den jeweils definierten Datentypen und der Syntax (der Aufbau der Zeichenkette) spielen für die maschinelle Datenverarbeitung eine wichtige Rolle.

Im Data-Mining-Kontext relevant ist die Unterscheidung zwischen (Abb. 2.3 rechts)

- *konkreten Datentypen*, die meist einzelne numerische Werte mit unterschiedlichen Mess- oder Skalenniveaus haben (bei der weiteren Differenzierung unterscheidet man u. a. zwischen stetigen Merkmalen mit metrischen Werten wie z. B. Längen- oder Gewichtsangaben sowie kategoriellen Merkmalen mit ordinalskalierten Werten [haben Rangfolgen wie z. B. Schulnoten oder Scores] oder nominalskalierten Werten [Klassen ohne Ränge, z. B. Blutgruppe]) und

- *abstrakten Datentypen*, die komplexere Strukturen statt einzelner Werte haben (bei der weiteren Differenzierung unterscheidet man u. a. Bilder, Videos, akustische Aufnahmen, Texte, Webseiten, Genomsequenzen).

Für konkrete Datentypen gibt es bereits seit langem vielfältige mathematisch-statistische Analyseverfahren (Kap. 2.3.2). Abstrakte Datentypen gewinnen seit einigen Jahren erheblich an Bedeutung. Oft wird datentechnisch Abstraktes in einem ersten Schritt mittels Datenlabeling und Codierung ganz oder zumindest sequenziell in Konkretes überführt und dann weiterverwendet (Kasten 2.1). Teilweise gibt es dafür bereits standardisierte Zuordnungs-, Berechnungs- und Darstellungsvorschriften. Zu den Vorreitern zählen u. a. Zeit- und Raumangaben (ISO 8601 für Zeitangaben und die Serie ISO 19100 durch die nicht nur Raumbezüge, sondern auch vielfältige Metadatenelemente definiert werden). Im Cholerabeispiel ist die Ortsangabe mit Straßenbezeichnung und Hausnummer an sich abstrakt, die Geoposition mit der definierten Syntax jedoch konkret (Abb. 2.3). Die maschinelle Analyse vielfältiger Sachverhalte wird durch die Überführung von abstrakten in konkrete Datentypen erheblich befördert. In Bezug auf das historische Beispiel des Choleraausbruchs in London können die Adressen der Todesopfer und die Standorte der Wasserbrunnen zwar seit langem in einer Datenbank als abstrakte Datentypen gespeichert werden, aber erst seitdem jedes Objekt numerisch georeferenziert werden kann und es digitale Oberflächenmodelle mit spezifischen Karten diverser infrastruktureller Komponenten und Netze gibt, auf denen diese Objekte verortet werden können, lassen sich Entfernungen auch maschinell berechnen und z. B. räumliche Cluster- und Hotspot-Analysen maschinell durchführen.

Kasten 2.1 Codierungen von Objekten und Merkmalen (DIN 6763)

Ein Code ist eine festgelegte Folge von Zeichen. Man unterscheidet:

- *Identifizierende Codes*: Sie sollen Objekte eindeutig identifizieren und werden deshalb möglichst nur einmal vergeben. Ortsbezogenen Merkmalen wird dieser eindeutige Code z. B. mittels Georeferenzierung zugewiesen. Identifizierende Codes für Personen (z. B. Versicherten- oder lebenslange Arztnummern; Kap. 4 u. 5) können auch zur Pseudonymisierung verwendet werden, wenn die jeweilige Referenzierungsvorschrift bzw. die Schlüsseltabelle nicht allgemein zugänglich sind.
- *Klassifizierende Codes*: Sie sollen Objekte anhand von Merkmalsausprägungen definierten Klassen oder Gruppen zuordnen. Sie können sehr grob einteilen (z. B. Gesundheitszustände gesund, auffällig, krank) oder hochdifferenziert systematisieren (z. B. die internationale Krankheits-

klassifikation ICD[7] mit inzwischen mehr als 14.000 Kategorien). Dafür gibt es unterschiedliche Klassifikationsverfahren, z. B.:
  - Scoringverfahren weisen Objekten/Sachverhalten anhand definierter Merkmalsausprägungen Klassen mit Rangfolgen zu, damit kann ein Zustand (z. B. Krankheit) beschrieben aber auch eine Prognose abgeleitet werden (z. B. Risiken).
  - Objekterkennungsverfahren identifizieren Objekte anhand von Merkmalsausprägungen und ordnen sie Klassen/Features zu (z. B. Zeichen/Worte in der Texterkennung; Personen/Gewebsstrukturen in der Bilderkennung; Bewegungen/Situationen bei der Videoüberwachung).
  - Indexieralgorithmen sind regel- oder wahrscheinlichkeitsbasierte Zuordnungen von Objekten/Sachverhalten zu komplexeren Systematiken (z. B. Zuweisung von Vergütungspauschalen in Krankenhäusern; Kap. 4.4.1).
- *Mischung* aus identifizierenden und klassifizierenden Codes sollen neben der eindeutigen Identifikation zusätzlich einzelne Merkmalsausprägungen mitcodieren.

2.2.2 Datenspeicherung und -bereitstellung: von Datenbanken bis Systemarchitekturen

Für heutige Data-Mining-Verfahren relevante Daten werden in Datenbanken gespeichert, die unterschiedliche Formen und Strukturen haben können. Tabellen sind die klassische Form, um Daten objektbezogen strukturiert zu speichern (Abb. 2.3).

Mit *relationalen Datenbanken* können vielfältige Daten in diversen, im Vorfeld definierten Tabellen strukturiert abgelegt und über Schlüsselmerkmale bzw. deren Werte in Relation zu anderen Tabellen und deren Daten gesetzt werden. Dadurch lassen sich vielfältige Bezüge zu anderen Objekten und/oder Sachverhalten herstellen. Relationale Datenbankmodelle sind vor allem beim Umgang mit konkreten Datentypen gebräuchlich. Aufgrund der weiten Verbreitung und langjährigen Dominanz dieser Datenbankmodelle gibt es vielfältige Algorithmen, die speziell zur Analyse derart merkmalsbetont strukturierter Datenbestände entwickelt wurden. Diese Algorithmen können als Funktionalitäten in die Datenbank- und/oder Datenanalysesoftware integriert werden und sind dadurch leicht einsetzbar. Relationale Datenbanken kommen im Umgang mit

7 Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme (International Statistical Classification of Diseases and Related Health Problems, derzeit in der 10. Version – ICD-10) ist ein weltweit anerkanntes Klassifikationssystem für die Diagnose von Krankheiten. Sie wird von der Weltgesundheitsorganisation (World Health Organisation – WHO) herausgegeben und weiterentwickelt (ab 2022 in 11. Version).

komplex strukturierten Objekten/Sachverhalten an ihre Grenzen (TAB 2014, S. 46 ff.). An diesen Grenzen werden sie zunehmend durch nichtrelationale Datenbankmodelle ergänzt und erweitert, z. B. objekt- oder dokumentenorientierte Datenbankmodelle oder Mischformen.

Objektorientierte Datenbanken strukturieren Dateninhalte statt in zweidimensionalen Tabellen in mehrdimensionalen Datentensoren (vielschichtige mathematische Konstrukte). Einzelne Datenobjekte werden nach wie vor anhand von Merkmalen/Attributen beschrieben (vorzugsweise mittels konkreter Datentypen), können über Klassen/Features jedoch besser strukturiert und klassifiziert werden. Objektorientierte Datenbanken erlauben komplexere Datenstrukturen, die sich bei großen Datenmengen effizienter maschinell verwalten und verarbeiten lassen, auch wenn sie kaum anschaulich dargestellt werden können (ausführlicher z. B. in Bernsdorf et al. 2015, S. 30 f.). Objektorientierte Datenbanken sind gegenwärtig noch nicht so weit verbreitet wie relationale Datenbanken. Deren Analysetools sind tendenziell weniger vielfältig und rechenintensiver als die von zweidimensionalen Matrizen/Tabellen.

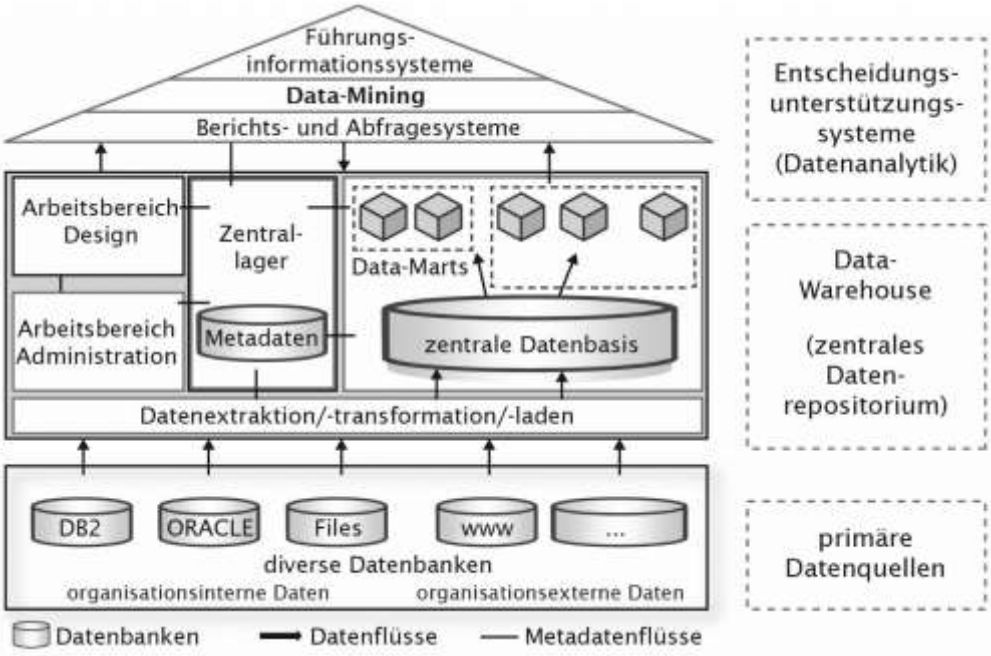
Bei *dokumentenorientierten Datenbanken* ist jeder Eintrag ein eigenes Dokument. Sie sind besonders geeignet für abstrakte Datentypen (z. B. Genomsequenzen, Texte, Bilder, Videos), die oft nur semistrukturiert vorliegen. Über eindeutige Identifikatoren (Objektstammdaten, Schlüsselmerkmale, Datenlabel) ist eine Verknüpfung unterschiedlicher Datenbankmodelle und der darin gehaltenen Datenbestände möglich. Derartig vielfältig verknüpfbare Bestände werden auch als poly- oder heterogen strukturiert bezeichnet.

Die Möglichkeit, unterschiedliche Datenbestände auf der Grundlage von gemeinsamen Standards und Schlüsseln miteinander zu verknüpfen, wird mit dem Begriff *Interoperabilität* assoziiert. Interoperabilität hat zwei wesentliche Ebenen, eine semantische (Verwendung einheitlicher Codierungen in Bezug auf die die Dateninhalte) und eine syntaktische (Nutzung einheitlicher Darstellungsstrukturen, Formate, Metadaten). Der Geodatenbereich mit seiner Normungsserie ISO 19100 kann als ein Vorreiter in Bezug auf interoperable Datenstrukturen aufgefasst werden. In anderen Bereichen (z. B. in medizinischen Bereich) sind interoperable Datenstrukturen noch eine große Herausforderung (Kasten 4.2).

Für die dauerhafte Speicherung und Mehrfachnutzung großer polystrukturierter Datenbestände gibt es unterschiedliche Systemarchitekturen: *Data Warehouses* in einrichtungsinternen Rechenzentren in Eigenverantwortung dieser Einrichtung (Abb. 2.4) und *Cloudstrukturen* mit mehreren vernetzten Rechenzentren, deren Hard- und Softwarekomponenten teilweise geöffnet und temporär vermietet werden. Beide Systeme können ihre Datenbestände nur dann effizient verwalten, wenn alle enthaltenen Daten strukturiert durch Metadaten(files) beschrieben werden.

Data Warehouses sind zentrale, permanente Datenrepositorien einer Einrichtung, in die potenziell relevante Daten aus diversen heterogenen Primärdatenbanken regelmäßig physisch zusammengeführt, geprüft, konsolidiert, verdichtet, in definierte Formate transformiert und dauerhaft gespeichert werden (Bernsdorf et al. 2015, S. 36; Rahm 2015; Schepers et al. 2015, S. 130 ff.; TAB 2014, S. 43). Zwar können administrative Prozesse des Datenzugangs und der Datenverwaltung zentral organisiert werden, dennoch sind Aufbau, kontinuierliche Datenkonsolidierung, Betrieb und Weiterentwicklung mit erheblichem Aufwand verbunden. Je geringer der Interoperabilitätsgrad primärer Datenquellen ist, desto aufwendiger ist der Transfer. Dieser Aufwand lohnt sich nur, wenn eine Weiterverwendung der Daten explizit anvisiert wird. Für jede Weiterverwendung werden dann spezifische Auszüge erstellt (Data-Marts) (Abb. 2.4). Fallweise können dabei spezifische Datenschutzmaßnahmen (z. B. Anonymisierungen und/oder Vergröberungen) vorgenommen werden, ohne die zentrale Datenbasis zu verändern.

Abb. 2.4 Grobarchitektur von Datenhaltung und Analyse



Quelle: Rahm 2015, S. 11

Auch die datenanalysierenden Softwarekomponenten werden in der Regel fachlich kontrolliert, genutzt und weiterentwickelt. Sie halten diverse anwendungsbezogene Analysewerkzeuge bereit, mit denen z. T. nach neuen Strukturen und Mustern in den Daten gesucht werden kann (Schepers et al. 2015, S. 183 ff.). Gegebenenfalls lassen sich einzelne Verfahren oder mathematisch-statistische

Modelle operationalisieren und z. B. zu Berichts-, Abfrage- oder Informationssystemen verstetigen (z. B. Wetterdienste). *Informationssysteme* stellen das Informationsbedürfnis von Anwendenden in den Mittelpunkt und bieten Services an, die aus maschinenlesbaren Daten für Menschen unmittelbar erfassbare Informationen extrahieren (soziotechnische Perspektive). Beispielsweise können mit spezifischen Geoinformationssystemen (GIS) sowohl Wege zwischen zwei Punkten mathematisch optimiert werden als auch ermittelte Ergebnisse auf den digitalen Landkarten visualisiert und damit leichter erfasst werden.

Bei Data Warehouses ist der Zugang zu den Daten und den Analysetools sowie zu den damit generierten Ergebnissen und Informationen kontrolliert und begrenzt. Sie werden u. a. in der medizinischen Forschung und im Gesundheitssystem genutzt (Fallstudien in Kap. 4 u. 5).[8] Data Warehouses kommen beim Umgang mit sehr großen Datenbeständen und bei Analysen, die möglichst zeitnah zur Datenerhebung realisiert werden sollen, an ihre Grenzen. Denn erst müssen Daten in das Repositorium integriert werden und dann werden die Daten zu den Auswertesystemen transferiert und nicht umgekehrt. Insbesondere wenn Festplatten, die zu den vergleichsweisen langsamen Bestandteilen der Rechnerarchitektur gehören, als Datenspeicher fungieren, werden Datenabfragen und -analysen relativ zeitaufwendig. Mehrere hardwaretechnologische Entwicklungen befördern die schnelle Analyse immer größerer Datenbestände.[9] Dennoch eignen sich derartige permanente Datenrepositorien nicht für Echtzeit- oder Nahe-Echtzeitanalysen von Daten, die nur temporär relevant sind (z. B. zur Automatisierung unterschiedlicher Arbeitsprozesse im Rahmen von Industrie 4.0 oder des autonomen Fahrens).

Die Möglichkeiten des massiv-parallelen Rechnens mit verteilt arbeitenden Computerprogrammen und die Nutzung von Computerclustern, die über Serverknoten sowohl die Datenspeicherung dezentralisieren als auch die Rechenleistung erweitern (skalieren) können, führen zu einer neuen Systemarchitektur – des *Cloudcomputing* (ausführlich z. B. in TAB 2014). Dafür muss die Hardware nicht mehr lokal in einem Rechenzentrum stehen, sondern kann bedarfsorientiert und flexibel über ein Netzwerk eingebunden werden. Daten werden temporär oder permanent auf unterschiedlichen Servern verteilt gelagert

8 Der Begriff Data Warehouse wird wegen der Warenhausassoziation teilweise abgelehnt, vor allem wenn datenhaltende Stellen hochsensible Daten sicher verwahren, keinesfalls weitergeben und vor dem Zugriff unberechtigter Dritter in besonderem Maße schützen. Das Nationale Centrum für Tumorerkrankungen (NCT) bezeichnet sein Repositorium genetischer Daten als DataThereHouse (www.nct-heidelberg.de/forschung/nct-core-services/nct-datatherehouse.html; 13.12.2021).

9 Durch die In-Memory-Technologie können die schnelleren, zunehmend großen Arbeitsspeicher als Datenspeicher genutzt und langsamere Festplattenoperationen vermindert werden (TAB 2014, S. 48). Der Einsatz von Multi- oder Many-Core Prozessoren, von Grafikprozessoren bzw. -karten oder von spezifischen Chips wie Tensorprozessoren ermöglichen schnelles und paralleles Ausführen von Rechenoperationen (DEK 2019, S. 63). Damit steigt die Leistungsfähigkeit von Data Warehouses erheblich.

und dann unter Minimierung des Datentransfers dezentral und parallel verarbeitet. Die sich über viele miteinander vernetzte Computer erstreckende Datenverarbeitung erfordert flexible Strukturen und Interoperabilität aller beteiligten Komponenten. Zentrale Elemente sind Verteilsysteme, die Daten und Rechenoperationen über ein Netzwerk unterschiedlicher Rechner verteilt verwalten und parallel ausführen können (ausführlicher z. B. Bernsdorf et al. 2015, S. 162 ff.). Dadurch entstehen unterschiedliche Cloudservicebereiche. Sie ermöglichen die flexible Bereitstellung und Nutzung von:

- *Hardwareressourcen* (Infrastructure as a Service) mit seinen Komponenten Rechner, Netze und Speicher;
- *Programmierumgebungen* (Platform as a Service), um neue datenanalytische Funktionalitäten und Anwendungen leicht entwickeln und Serviceangebote kontinuierlich erweitern zu können;
- *Analysesoftware* (Software as a Service), wobei Anbietende ihre Software auf ihren Computerclustern betreiben und Clienten erlauben, diese ebenfalls zu nutzen;
- *Daten* (Data as a Service), können nicht nur aufbewahrt, sondern auch geteilt und genutzt werden, wobei sich unterschiedliche Möglichkeiten eröffnen, je nachdem ob lediglich Datenstrukturen, inhaltlich relevantere Rand-/Verkehrs-/Metadaten oder auch Dateninhalte zugänglich gemacht werden;
- *Sicherheitskonzepten* (Security as a Service), für alle Elemente der Datenverarbeitung können das Identitätsmanagement (Autorisierung, Authentifizierung), die Datenverschlüsselung, Sicherheitsüberwachungen sowie Wartung und Aktualisierung von Sicherheitsprogrammen übernommen werden;
- sodass durch diese unterschiedlichen Serviceelemente zum Teil ganze *Geschäftsprozesse* (Business Process as a Service) über Cloudstrukturen abgewickelt werden können.

Cloudcomputing benötigt zentrale Steuerungseinheiten, die diverse parallellaufende Transaktionen und Aktivitäten orchestrieren.[10] Auch wenn viele Datenanalysen ohne Cloudstrukturen durchgeführt werden können und Cloudcomputing zahlreiche Anwendungsfelder auch jenseits der Datenanalytik hat, wird insbesondere bei der sehr schnellen Analyse großer Datenmengen Cloudcomputing ein besonderes Potenzial zugeschrieben. Data-Mining-Prozesse mit rechenintensiven Ansätzen wie beispielsweise das Training künstlicher neuronaler Netze (Kap. 2.3.2) benötigen Datenverarbeitungsleistungen, die bisher am ehesten durch Cloudstrukturen erreicht werden können (Anwendungsbeispiele in Kap. 4.3.3).

10 Das TAB hat bisher keine Hinweise, die darauf hindeuten, dass gänzlich andere Architekturen mit dezentral autonomen Organisationseinheiten ohne zentrale Instanz (z. B. dezentrale Peer-to-Peer-Netzwerke oder Blockchainkonzepte) im Kontext von Data-Mining als Knowledge Discovery in Databases von besonderer Relevanz sind.

Cloudservices werden mit diversen Geschäftsmodellen betrieben. Die Spanne reicht von privaten Clouds, die von einem Unternehmen selbst in eigenen Rechenzentren betrieben und ausschließlich firmenintern genutzt werden, bis zu öffentlichen Clouds mit weltweit verteilten Rechenzentren, die Speicher- und Rechenleistungen skalieren und hochflexibel bereitstellen können. In private Clouds können wie in Data Warehouses diverse externe Datenbestände integriert werden, deren Weiterverwendung ist jedoch begrenzt und kontrolliert. Bei öffentlichen Clouds können Nutzende kaum Einfluss darauf nehmen, wie und wo ihre Daten gespeichert und verarbeitet werden. Die größten öffentlichen Clouds werden teilweise auch als Hyperscaler[11] bezeichnet.

Besonderes Potenzial wird Systemarchitekturen unterstellt, bei denen Betreiberorganisationen zum einen Wert legen auf interoperable Datenkomponenten und auf einen effizienten Zugang zu Datenbeständen Dritter und zum anderen attraktive Programmierumgebungen und Angebote für Nutzende schaffen. Datenkomponenten benötigen Kommunikationskanäle, über die datengebende und datenverarbeitende Akteure eine Datennutzung situativ vereinbaren und rechtskonform realisieren können. In einer begrifflichen Erweiterung werden teilweise *Dateninfrastrukturen* gefordert. Durch diese sollen vor allem Daten, die im Rahmen öffentlicher Aufgaben entstanden, aus vielfältigen Quellen rechtskonform zugänglich und maschinell nutzbar gemacht werden. Das erfordert Harmonisierung und Katalogisierung (standardisierte Metadatensätze) von Datenbeständen sowie die Vernetzung von Repositorien und rechtssichere Zugänge über Portale. Öffentliche Geodatenbereiche gelten diesbezüglich als Vorreiter (Kap. 3.2.). Wenn dazu Betreiber attraktive Programmierumgebungen schaffen, sodass viele Softwareentwickler neue datenbezogene Anwendungen leicht erarbeiten und anbieten können, mit denen wiederum die Zahl der Nutzenden steigt, die darüber immer vielfältigere wirtschaftliche und private Prozesse abwickeln können, bezeichnet man dies teilweise auch als *digitale Ökosysteme* (Kasten 2.2). Ansätze, die nicht nur im Geschäftskundenbereich agieren (Business-to-Business), sondern auch Privatpersonen/Verbraucher/Patienten ansprechen können (Business-to-Consumer), wird besonderes Potenzial auch für gesellschaftliche Veränderungen unterstellt.

11 Hyperscaler verbinden Tausende bis ggf. Millionen Server in einem Netzwerk. Dadurch werden hohe Zugriffsraten und fluktuierende Nutzung gleichermaßen ermöglicht, man spricht von horizontaler Leistungsskalierung oder auch von Grid-Computing (TAB 2014, S. 57). Vor allem die großen Clouds von Amazon, IBM, Google und Microsoft gelten als Hyperscaler.

Kasten 2.2 GAIA-X

GAIA-X ist eine maßgeblich durch deutsche Akteure aus Politik, Wirtschaft und Wissenschaft vorangetriebene Initiative, ein europaweit vernetztes digitales Ökosystem aufzubauen (BMWi 2019). Ziel ist es, ein modulares, sicheres, vertrauenswürdiges und nutzerfreundliches Verbundsystem von europäischen Anbietern unterschiedlicher Cloudservices zu schaffen, das europäische Vorgaben zu Rechts-, Daten- und Cybersicherheit technisch umsetzen kann und einen souveränen Umgang mit Daten und Anwendungen (Analysetools) gewährleistet. Laut Bundesregierung (2020b) handelt es sich um ein neues Konzept einer vernetzten Dateninfrastruktur, das weitgehend auf bereits vorhandenen Elementen aufbauen und diese über offene Schnittstellen und Standards verbinden soll (kein gänzlich neuer Hyperscaler). Ziel sei ein gemeinsames digitales Ökosystem von Anbietenden und Anwendenden aus Wirtschaft, Wissenschaft und öffentlichen Diensten. Unter anderem sollen die über die nationale Forschungsdateninfrastruktur (NFDI) vernetzten Forschungsdatenzentren (S. 119) an diese IT-Architektur angeschlossen werden.[12] Mit GAIA-X wird die Hoffnung verknüpft, dass damit interoperable Datenstrukturen entstehen, durch die auch im nationalen Gesundheitssystem Hunderttausende Lerndatensätze in geprüfter Qualität für vielfältige Data-Mining-Projekte rechtskonform bereitgestellt werden können (BMWi 2019, S. 27 ff.). Eine prototypische Implementierung der Basisfunktionalität war ursprünglich für Ende 2020 geplant. Danach müsse sich GAIA-X am Markt beweisen. Die Bundesregierung kündigte an, GAIA-X als Kernelement der souveränen Datennutzung entschieden voranzutreiben (Bundesregierung 2021a, S. 21) und fördert inzwischen vielfältige Anwendungsentwicklungen auch im Gesundheitsbereich.[13]

Data Warehouses und Cloudcomputing haben etliche Gemeinsamkeiten aber auch jeweils spezifische Stärken und Grenzen, beide haben datenhaltende und datenanalysierende Komponenten. Cloudstrukturen können durch die Vernetzung von Hard- und Softwarekomponenten je nach situativem Bedarf flexibel genutzt werden. Durch standardisierte Datenkomponenten und Programmierumgebungen können viele softwareentwickelnde Akteure eingebunden werden und lassen sich datenanalytische Werkzeuge tendenziell schneller weiterentwickeln. Je offener Cloudstrukturen sind, desto aufwendiger und schwerer ist die Kontrolle der ablaufenden Prozesse sowie die rechtliche Klärung von Verantwortungs- und Haftungsfragen. In Data Warehouses lassen sich datenverarbei-

12 www.nfdi.de/fair-data-spaces/ (13.05.2022)
13 www.bmwi.de/Redaktion/DE/Dossier/gaia-x.html (13.12.2021)

tende Prozesse tendenziell besser kontrollieren und Verantwortlichkeiten rechtlich einfacher zuweisen. Data Warehouses und Cloudcomputing können sich ergänzen. Daten mit besonderer Kritikalität werden oftmals in Data Warehouses gehalten. Offene Cloudstrukturen werden für die Verarbeitung weniger kritischer Daten oder für spezielle Berechnungen genutzt, wenn eigene Ressourcen temporär nicht ausreichen.

Sicherheit von Systemkomponenten

Um die Anwendungsmöglichkeiten von Data-Mining zu erweitern sowie das Vertrauen in und die Akzeptanz von entsprechenden Prozessen zu stärken, sind sichere und verlässlich funktionierende Hard- und Softwarekomponenten erforderlich (Müller-Quade/ et al. 2020). Spezifische Sicherheitskonzepte nehmen die gesamte Prozesskette von der Primärdatenerhebung und deren Bereitstellung als Trainingsdaten über die Analyse und Ableitung von Informationen bis hin zur Operationalisierung datenanalytischer Verfahren und der Entwicklung von Informationsdiensten in den Blick und versuchen, alle Komponenten zuverlässig und unter Einhaltung normativer Vorgaben bereitzustellen. Derartige Sicherheitskonzepte haben mehrere Teilbereiche mit jeweils spezifischer Ausrichtung: IT-Sicherheit, Datenschutz und -sicherheit sowie Sicherheit/Richtigkeit im Rahmen des analytischen Vorgehens. IT-Sicherheitskonzepte sind auf die zuverlässige Bereitstellung von Hard- und Softwarekomponenten ausgerichtet und zielen darauf ab, Störungen möglichst prospektiv zu verhindern. Störungen werden in der heutigen Praxis wesentlich durch Cyberattacken verursacht. Sie können auch Data-Mining-Aktivitäten auf unterschiedliche Art und Weise gefährden, u. a. wenn

- Roh- und/oder Trainingsdaten gelöscht oder unzugänglich gemacht werden und in Folge Data-Mining-Prozesse nicht mehr realisiert, reproduziert oder geprüft werden können;
- Trainingsdaten manipuliert werden (teilweise können bereits geringfügige Datenmanipulationen Data-Mining-Ergebnisse verändern; Kap. 2.3.3) oder
- Analyseverfahren oder resultierende Ergebnisse manipuliert oder blockiert werden.

Die Bewertung der Gefährdungslage von IT-Systemkomponenten erfolgt situativ und soll mögliche Folgedimensionen prospektiv in den Blick nehmen. Dabei wird vor allem rechtlich differenziert nach der gesellschaftlichen Bedeutung der jeweiligen Anwendungsbereiche (Einstufung als Komponenten kritischer Infrastrukturen; Kasten 3.1), aber auch nach den jeweiligen Dateninhalten und Geheimhaltungsinteressen. IT-Sicherheits- und Datenschutzkonzepte (Kap. 3.3.3)

sind sich ergänzende notwendige Elemente für vielfältige Data-Mining-Prozesse.

Die Sicherheit von IT-Systemen kann mit unterschiedlichen Ansätzen angestrebt werden. Der klassische Ansatz versucht, alle Komponenten mit unterschiedlichen technischen, organisatorischen und personellen Maßnahmen präventiv vor Angriffen zu schützen. Dazu gehören u. a. Datenverschlüsselungstechniken, der Einsatz von Firewall-Software, Zertifizierungs- und Autorisierungsverfahren, Protokollierungen und kontinuierliche Systemüberwachungen bezüglich unberechtigter Zu- und Eingriffe. Die Expertenkommission Forschung und Innovation (EFI) fordert, der Thematik Cybersicherheit mehr Bedeutung beizumessen und u. a. diesbezügliche nationale Kompetenzen auszubauen, Qualität entsprechender Produkte und Dienstleistungen zu verbessern sowie bestehende Standards und Zertifizierungen weiterzuentwickeln (ausführlich EFI 2020, S. 42 ff.).

IT-Sicherheitsexpert/innen bezweifeln jedoch, dass es in absehbarer Zeit möglich sein wird, vernetzte Systemkomponenten vollständig sicher auszugestalten. Sie empfehlen deshalb zusätzlich zum klassischen IT-Sicherheitsansatz, auch mögliche Folgen solcher Attacken zu antizipieren und zu versuchen, diese möglichst gering zu halten. Die Datenethikkommission betont in diesem Zusammenhang auch die Bedeutung des Erhalts menschlicher Kompetenzen und Kontrollfähigkeiten (DEK 2019, S. 165).

Die IT-Sicherheitsperspektive ist darauf ausgerichtet, Angriffe von außen abzuwehren und damit möglicherweise verbundene Schäden möglichst zu minimieren. Die Richtigkeit des analytischen Vorgehens, der verwendeten Roh- und Analysedaten sowie der jeweiligen mathematischen Verfahren und Algorithmen und die Zuverlässigkeit der eingesetzten Analysesoftware wird dabei nicht hinterfragt. Derartige Prüfungen sollten Bestandteil des eigentlichen Data-Mining-Prozesses (Kap. 2.3). bzw. in bestimmten Anwendungskontexten wie z. B. in der Medizin auch von Prüfinstanzen sein.

2.3 Data-Mining als Prozess: Schritte, Verfahren, Ergebnisse

2.3.1 Spezifikation der Untersuchungsaufgabe und Datenaufbereitung

Data-Mining-Prozesse sind sowohl daten- als auch anwendungsgetrieben, d. h., es gibt zum einen Datenbestände und zum anderen spezifizierbare Untersuchungsaufgaben oder konkrete Fragestellungen, wie z. B.: Wo sind potenzielle Zielgruppen für ein Produkt oder eine Kampagne? Was sind relevante Risikofaktoren für Krebs? Welche Datenmuster sind spezifisch für bestimmte Sachverhalte oder Objekte? Bei einer datengetriebenen Aufgabenbearbeitung wird

jede Untersuchungsfrage dann als ein mathematisches Problem formuliert, um dieses möglichst anhand verfügbarer Datenbestände mit einem bereits existierenden Analyseverfahren zu lösen (Knobloch/Weidner 2000; Schepers et al. 2015, S. 43; Zweig 2016). Aus analytischer Sicht werden mit Data-Mining folgende Problemkategorien und Verfahren in Verbindung gebracht:

- *Erkennen von Ähnlichkeiten/Clusteranalysen*: Verfahren suchen nach ähnlichen Objekten und nach Ähnlichkeitsstrukturen. Ziel ist es, Ballungen/Häufungen in Datensätzen zu erkennen und Gruppen ähnlicher Objekte zu identifizieren (Klassen/Features erstmals bilden), ohne dass Vorwissen berücksichtigt wird (z. B. bereits bestehende Klassifikationen).
- *Erkennen von Anomalien/Ausreißern*: Verfahren suchen Auffälligkeiten in Datenbeständen, die auf Besonderheiten oder auf mögliche Datenfehler hindeuten und genauere Untersuchungen erfordern.
- *Objektzuweisungen/Klassifizierung*: Verfahren ordnen neue Objekte bereits bestehenden Klassen oder Gruppen zu.
- *Erkennen von häufigen Zusammenhängen/Assoziationsanalysen*: Verfahren suchen nach Beziehungen zwischen unterschiedlichen Merkmalen/Variablen (Korrelationen) und leiten daraus Regeln ab (welche Merkmalsausprägungen treten häufig zusammen auf, z. B. beim Kaufverhalten).
- *Erkennen von Strukturen zwischen Merkmalen und deren Ausprägungen/Regressionsanalysen*: Verfahren quantifizieren Beziehungen vorzugsweise zwischen metrischen Merkmalen/Variablen. Die Resultate (Formeln, Modelle) können im nächsten Schritt zur Prognose fehlender Werte eingesetzt werden.
- *Reduktion der Datenmenge/Zusammenfassungen* (nutzen alle genannten Verfahren): Verfahren zielen darauf ab, repräsentative Datenteilmengen zu finden, die möglichst nur die relevanten Teile einer Gesamtheit enthalten. Sie sind vor allem bei abstrakten Datentypen mit komplexen Strukturen (Texte, Bilder, Videos) relevant, um Trainingsdatensätze zu ermitteln (Extraktion) oder Schlüsselsequenzen, -begriffe abzuleiten (Abstraktion), anhand derer im nächsten Schritt z. B. Suchmaschinen[14] optimiert werden.

Da für unterschiedliche Datentypen unterschiedliche Verfahren in Betracht kommen und jedes Verfahren Stärken und Grenzen hat, sind bei der Verfahrensfestlegung erhebliche Fachkenntnisse erforderlich. Auch muss sich bereits abzeichnen, welche Rohdatenbestände genutzt werden sollen. Fehler bei der Spezifikation des methodischen Vorgehens können Ergebnisse und damit den ganzen Data-Mining-Prozess unbrauchbar machen.

14 Suchmaschinen sind Programme zur Recherche von Begriffen und Dokumenten, um Anfragen gezielt zu beantworten und Ergebnisse in einer möglichst sinnvollen Form bereitzustellen.

Aus verfügbaren Datenbeständen werden entsprechend der Untersuchungsaufgabe erforderliche Analyse- bzw. Trainingsdatensätze definiert, extrahiert und aufbereitet. Der Aufwand für die Datenaufbereitung in einem Data-Mining-Projekt ist oftmals erheblich. Er liegt Schätzungen zufolge bei 50 bis 80 % der Projektressourcen (Knobloch/Weidner 2000, S. 354; Phillips 2017, S. 731). Die notwendigen Datenaufbereitungsschritte werden in der Regel projektspezifisch definiert und realisiert, wobei u. a. Verfahrensregeln und Filter kontinuierlich geprüft und weiterentwickelt werden (anwendungsbezogene Darstellung in Kap. 4 u. 5). Um relevante Auffälligkeiten und Anomalien in Daten von Fehlern unterscheiden zu können, sind vielfältige Kenntnisse zum Entstehungsprozess der Daten, zu den abgebildeten Sachverhalten und den jeweiligen Kontexten erforderlich. Dafür erforderliche Arbeitsschritte lassen sich bisher weit weniger standardisieren oder gar automatisieren als die eigentliche Datenanalyse. Weitgehend übereinstimmend warnen Experten vor einem bedingungslosen Vertrauen in die Richtigkeit der Inhalte großer Datenbestände und dem blinden Einsatz von Data-Mining-Techniken. Welches Maß an Richtigkeit ein Analysedatensatz haben muss bzw. welches Maß an Fehlern für eine Datenanalyse toleriert werden kann, lässt sich oft nur situativ entscheiden.

2.3.2 Datenanalytische Verfahren

Data-Mining zielt darauf ab, in Datenbeständen enthaltene strukturelle Muster aufgabenbezogen zu ermitteln und darzustellen. Diese Darstellungen können sehr unterschiedlich sein. Die Spanne reicht von einzelnen statistischen Koeffizienten oder Ähnlichkeitsmaßen über die Ableitung von Regeln und Zuordnungsvorschriften bis zur Anpassung und Parametrisierung von Funktionen – auch als mathematisch-statistische Modelle bezeichnet. Die zunehmende Datenerhebung öffnet in vielfältigen Lebensbereichen Türen für die mathematische Modellierung, die versucht, wesentliche Phänomene eines Sachverhalts mittels Formeln zu beschreiben. Beim datenbasierten Vorgehen werden in einer *Trainingsphase* mit unterschiedlichen analytischen Verfahren aufgabenbezogen anhand von (Trainings-)Daten strukturelle Besonderheiten und Muster gesucht, Entscheidungsregeln aufgestellt oder Parameter allgemeiner mathematischer Modelle an diese Trainingsdaten angepasst. Mathematische Modelle können trivial einfach sein (z. B. einfache logische Verknüpfungen und Entscheidungsregeln) oder hochkomplex (wenn Modelle viele Variablen mit nichtlinearen Beziehungen abbilden und viele Parameter anhand umfangreicher Datenbestände angepasst werden z. B. Wettermodelle). Diese Darstellungen der ermittelten strukturellen Muster sind das primäre Data-Mining-Ergebnis. Sie ermöglichen eine spezifische Auseinandersetzung mit den jeweils ermittelten Mustern und Beziehungen sowie deren Prüfung. Dabei wird untersucht, inwiefern ermittelte Strukturen und Modelle verlässlich, valide und generalisierbar sind, d. h. sie ggf. zu prognostischen Zwecken auf neue Situationen übertragen werden können.

Diverse klassische mathematisch-statistische Verfahren und Analysetechniken führen zu strukturellen Beschreibungen (logische Entscheidungsregeln, parametrisierte Formeln und Modelle), die zumindest für Fachleute nachvollziehbar und verständlich sind. Witten et al. (2011, S. 5 ff.) sind der Ansicht, dass derartige Strukturbeschreibungen in vielen Fällen mindestens so wichtig seien, wie deren möglicher Einsatz in neuen Situationen. Denn Data-Mining-Ergebnisse würden nicht nur genutzt, um neue Situationen zu bewerten, sondern wesentlich auch, um die ermittelten strukturellen Muster verstehen und Klassifikationsprozesse oder Vorhersagen nachvollziehen zu können – um im Sinne eines Knowledge Discovery in Databases auch Erkenntnisse abzuleiten bzw. Wissen zu erweitern. Deshalb sei es wichtig, in welcher Form Ergebnisse einer datenbasierten Trainingsphase präsentiert werden, entweder

- als nachvollziehbares mathematisches Modell, dessen Entscheidungsstruktur anhand von Formeln explizit dargestellt wird (symbolische Ansätze) oder
- als Blackbox, die weder relevante Merkmale ausweist noch Entscheidungsstrukturen nachvollziehbar darstellt (nichtsymbolische Ansätze).

Dieser Ansatz von Witten et al. (2011) erlaubt es, die Begriffe Data-Mining, maschinelles Lernen[15] sowie klassische mathematisch-statistische Verfahren und künstliche neuronale Netze (KNN) einzuordnen. Witten et al. interpretieren Data-Mining als anwendungsgetriebenes maschinelles Lernen und klassische mathematisch-statistische Verfahren sowie das Training künstlicher neuronaler Netze als sich ergänzende methodische Vorgehensweisen.

Maschinelles Lernen

Maschinelles Lernen steht für den Aufbau und die schrittweise Anpassung mathematisch-statistischer Modelle an einen Trainingsdatensatz, die nach der Trainings- und Validierungsphase zur Beurteilung neuer Fälle oder Situationen des gleichen Sachverhalts eingesetzt werden können. Dafür gibt es unterschiedliche Herangehensweisen. Grob unterscheidet man:

- *Überwachte Lernverfahren* beruhen auf Trainingsdatensätzen, in denen aufgabenspezifische Zielmerkmale explizit enthalten und kategorisiert sind. Anwendungsbereiche sind Klassifizierungen/Objekterkennungen oder die Entwicklung klassischer parametrisierter Modelle u. a. für prognostische Aufgaben wie z. B. Szenarioberechnungen.
- *Unüberwachte Lernverfahren* beruhen auf Trainingsdatensätzen, in denen aufgabenspezifische Zielmerkmale nicht explizit enthalten sind. Anwen-

15 Witten et al. (2011, S. 7 f.) plädieren dafür, statt Begriffe zu verwenden, die in erster Linie mit menschlichen Fähigkeiten assoziiert sind wie Intelligenz, Lernen oder Wissen und diese mit Adjektiven wie künstlich und maschinell abzugrenzen, nüchterner von komplexer Datenanalytik, von Trainingsprozessen und Modellanpassungen zu sprechen.

dungsbereiche sind Objekt- oder Merkmalsgruppieren (fassen ähnliche Objekte oder ähnliche Merkmale/Eigenschaften zusammen) oder Hauptkomponentenanalysen (strukturieren und vereinfachen umfangreiche Datensätze, fassen Merkmale zusammen).

Etliche klassische statistische Verfahren werden aktuell mit maschinellem Lernen assoziiert (z. B. Regressions-, Cluster-, Faktor- oder Hauptkomponentenanalysen). Sie parametrisieren mathematisch-statistische Modelle anhand eines (Trainings-)Datensatzes und führen zu Entscheidungsregeln für einen definierten Sachverhalt. Auch das Training künstlicher neuronaler Netze, um spezifische Sachverhalte zu erkennen und Aufgaben gezielt zu lösen, erfolgt mit speziellen, automatisiert ablaufenden (maschinellen) Lernverfahren bzw. -algorithmen. Man spricht vom Training bzw. der Anpassung eines allgemeinen statistischen Modells an einen Sachverhalt. Die Begriffe Data-Mining und maschinelles Lernen haben eine große Schnittmenge. Eine detaillierte Beschreibung unterschiedlicher datenanalytischer Werkzeuge und Verfahren, die mit maschinellem Lernen assoziiert werden, gibt z. B. Bishop (2006).

Klassische statistische Verfahren (symbolische Verfahren)

Viele Data-Mining-Aufgaben lassen sich mit klassischen multivariaten Verfahren lösen, insbesondere Bestände mit konkreten Datentypen können mit ihrer Hilfe vielfältig untersucht werden. Knobloch/Weidner (2000, S. 347) bezeichnen sie auch als nutzergeführte Verfahren, da Nutzende im Rahmen eines Data-Mining-Prozesses definieren, anhand welcher Merkmals-/Objektbereiche und mit welchen Verfahren und Algorithmen Untersuchungsaufgaben gelöst werden. Je nach Datentyp kommen unterschiedliche Algorithmen in Betracht, oder andersherum haben einzelne Algorithmen meist bestimmte Voraussetzungen und Annahmen in Bezug auf die zu analysierenden Daten (z. B., dass die Werte eines Merkmals normalverteilt sind, oder dass unterschiedliche Merkmale unabhängig voneinander sind). Wenn diese Annahmen nicht erfüllt und eingehalten werden, kann man sich nicht darauf verlassen, dass die ermittelte Struktur generalisierbar ist.

Teilweise werden klassische multivariate Verfahren auch als symbolische Verfahren bezeichnet, da die aus den Trainingsdaten ermittelten strukturellen Muster, Auffälligkeiten oder Zusammenhänge durch Kennziffern, Formeln oder Regeln explizit dargestellt werden. Beispiele für ermittelte Strukturen sind u. a. Checklisten oder Entscheidungsbäume zur Objektklassifikation oder prognostische Modelle (Anwendungsbeispiele in Kap. 4.3). Klassische multivariate Verfahren liefern neben den Strukturen in der Regel auch verfahrensspezifische Gütekriterien (z. B. Bestimmtheitsmaße, Homogenitätskoeffizienten, Signifikanzniveaus), anhand derer abgeschätzt werden kann, wie gut das ermittelte Ergebnis den Trainingsdatensatz repräsentiert (interne bzw. datenbasierte Prüfung in Kap. 2.3.3).

Bei komplexen Sachverhalten (die durch sehr viele Merkmale und/oder abstrakte Datentypen dargestellt werden) oder bei sehr selten auftretenden Ereignissen/Sachverhalten kommen klassische multivariate Verfahren an ihre Grenzen. Auch nach jahrelanger Weiterentwicklung liefern sie fehlerbehaftete oder unbrauchbare Ergebnisse, weil sie die Komplexität der jeweiligen Sachverhalte auch mit Tausenden von Regeln und Formeln nicht adäquat abbilden können. Beispiele sind Objekterkennungen auf Bildern oder Videos, Texterkennungen und -übersetzungen.

Auf unterschiedliche wahrscheinlichkeitstheoretische Ansätze aufbauende Verfahren können teilweise auch dann noch zuverlässige Ergebnisse liefern, wenn klassische multivariate Verfahren an ihre Grenzen kommen. Beispiele sind Fuzzylogiksysteme, die neben klaren Wahr- oder Falschaussagen bzw. eindeutigen Wenn-dann-Beziehungen auch unscharfe Aussagen zulassen (z. B. wenn X, dann zu 70 % Y), oder Bayes'sche Netze, die Wahrscheinlichkeitsmodelle faktorisieren und auch bei kleineren (Trainings-)Datensätzen eingesetzt werden oder auch bei sehr seltenen Ereignissen noch sinnvolle Ergebnisse liefern können (ausführlicher in Kap. 5.5.3). Derartige Ansätze können sinnbildlich als Brückenglied zwischen klassischen statistischen Verfahren und künstlichen neuronalen Netzen aufgefasst werden. Eine detaillierte Beschreibung unterschiedlicher datenanalytischer Werkzeuge und Verfahren, die beim Data-Mining eingesetzt werden können, geben z. B. Witten et al. (2011).

Training künstlicher neuronaler Netze (subsymbolische Verfahren)

Künstliche neuronale Netze können als eine Art allgemeines mathematisch-statistisches Modell mit höherer Komplexität aufgefasst werden. Deren Grundstruktur hat gewisse Ähnlichkeiten mit dem Aufbau des Gehirns, deshalb werden sie als künstliche neuronale Netze (KNN) bezeichnet oder mit dem Begriff künstlicher Intelligenz (KI) assoziiert. KNN bestehen aus künstlichen Neuronen/Knoten (es gibt unterschiedliche Formen), die auf Schichten (Layer) angeordnet und über gewichtete Verbindungen (mathematische Funktionen) miteinander verknüpft sind. Leistungsstarke KNN haben viele Knoten auf mehreren hintereinanderliegenden Schichten.[16] Auch die einzelnen Schichten sind über Aktivierungsfunktionen miteinander verbunden. Diese Funktionen können ebenfalls unterschiedliche Formen und formgebende Parameter haben (lineare, nichtlineare, auch mit Differenzialgleichungen wird experimentiert). Durch die jeweils eingesetzte Form der künstlichen Neuronen, die Anzahl der Schichten

16 2000 vernetzten KNN 102 Neuronen, 2015 bereits 106 Neuronen (Bitkom 2019, S. 12). Das 2012 von Google entwickelte Deep Convolutional Neural Network hatte 9 hintereinanderliegende Schichten. 2015 stellte Microsoft Research Asia das Project Oxford mit mehr als 150 Schichten vor. Sogenannte ResNets, mit denen biologische kognitive Prozesse abgebildet werden sollen, haben mehr als 1.000 Schichten (Gelitz 2019).

und die jeweilige Form der Verbindungsfunktionen mit ihren Gewichten und Parametern entstehen unterschiedliche Netzstrukturen (Topologien).

Auch KNN werden in der Trainingsphase an einen spezifischen Sachverhalt angepasst und trainiert, um eine bestimmte Aufgabe zu lösen (z. B. Objekte auf Abbildungen zu erkennen). Dafür werden überwachte oder unüberwachte Lernverfahren eingesetzt. Die einzelnen Schritte eines Lernverfahrens sind in Algorithmen definiert und verändern sich während des Lernprozesses nicht. Die Lernalgorithmen bauen auf unterschiedlichen wahrscheinlichkeitstheoretischen Ansätzen auf. Sie definieren, wie zuerst in der Trainingsphase über die Aktivierungsfunktionen die jeweiligen Aktivierungsschwellenwerte von künstlichen Neuronen angepasst, die Funktionen zwischen den Neuronen und zwischen den Netzschichten modifiziert (verstärkt oder abgeschwächt), die jeweils ermittelte Lösung der Aufgabe kontrolliert und die Richtig-/Falschbewertung als Feedback in den Trainingsprozess eingespeist werden. Dadurch wird eine Entscheidungsstruktur innerhalb eines KNN aufgebaut. Da diese Entscheidungsstruktur von den vorgelegten Daten und den spezifischen Lernalgorithmen beeinflusst wird – sie durch diese lernt –, wird dieser Ansatz immer mit maschinellem Lernen assoziiert.

KNN funktionieren mit großen Datenmengen oft besonders gut, denn die anhand von bedingten Wahrscheinlichkeiten ermittelten internen Entscheidungsstrukturen werden meist besser, je mehr Daten eingespeist werden. Deshalb sind KNN und deren Lernalgorithmen meist so aufgebaut, dass bei jedem neuen Datensatz die Gewichtung der simulierten Verbindungen zwischen den künstlichen Neuronen und die Parameter der Aktivierungsfunktionen der Netzschichten entsprechend der jeweiligen Lernregel angepasst werden können und KNN folglich auch in der Anwendungsphase kontinuierlich weitertrainiert werden (ausführlicher z. B. Angerer 2018; Bitkom 2019; Nielsen 2018; Rey/Wender 2018; Silver et al. 2017; Welzel/Grosch 2018.)

Das Training künstlicher neuronaler Netze wird mitunter auch als sub- oder nichtsymbolisches Verfahren aufgefasst, weil die erlernten Lösungswege und die Entscheidungsstrukturen nicht durch das sich mit jedem Trainingsschritt verändernde Modell explizit dargestellt werden und in Folge auch nicht jeder Schritt des Lernprozesses nachvollzogen werden kann. Datenanalyst/innen wählen anhand der jeweiligen Aufgabenstellung die Form des KNN und die Lernverfahren aus und bereiten die jeweiligen Trainingsdaten auf, ohne dass sie explizit definieren, anhand welcher Merkmals-/Datenbereiche Strukturen und Muster ermittelt werden sollen. Übereinstimmend wird darauf hingewiesen, dass die Qualität und Repräsentativität der jeweiligen Trainingsdatensätze von entscheidender Bedeutung sind (z. B. Ching et al. 2018; Jones 2014; Rey/Wender 2018; Wolfangel 2015) – genau wie bei klassischen statistischen Verfahren.

Künstlichen neuronalen Netzen werden gegenwärtig besondere Potenziale unterstellt. Mit ihnen werden große Hoffnungen geschürt (z. B., dass sie die Medizin individualisieren und optimale Behandlungsabläufe ermitteln können),

aber auch Ängste verbunden (z. B. vor nicht nachvollziehbaren Entscheidungen und unkontrollierbaren Robotern). Auch wenn sich viele Anwendungsvisionen zum Einsatz von KNN z. B. in der medizinischen Diagnostik, beim autonomen Fahren, bei der Technikwartung, Spracherkennung oder Identitätsfeststellung (ausführlicher z. B. in Bitkom 2015; Hecker et al. 2017) bisher kaum realisierten, hat durch große Forschungs- und Entwicklungsprogramme weltweit das diesbezügliche Technikverständnis zweifellos zugenommen und können Möglichkeiten und Grenzen realistischer eingeschätzt werden. Deutlich wird u. a., dass die jeweiligen Verfahren bisher nicht immer robust sind. Bereits geringfügige Datenveränderungen (z. B. durch technische Fehler bei der Datenerhebung, durch Verschleierungen oder Verzerrungen zur Erhöhung der Datensicherheit [Kap. 3.3.3] oder durch cyberkriminelle Aktivitäten) können Entscheidungsprozesse verändern. Diese Anfälligkeit erschwert den Übergang vom experimentellen KNN-Einsatz in die operative Anwendung (Ching et al. 2018; Finlayson et al. 2019; Heaven 2019). Aspekte der Produkt-, IT- und Cybersicherheit müssen beim KNN-Einsatz in besonderem Maße berücksichtigt werden (ausführlicher z. B. in Bitkom 2019, S. 43 ff.).

Trainierte künstliche neuronale Netze werfen auch Fragen zur Nachvollziehbarkeit der Ergebnisgenerierung und den Folgen auf. Welche Merkmale und Merkmalsausprägungen (Attribute) für die Unterscheidung und Zuordnung in die jeweiligen Zielkategorien entscheidungsrelevant sind und welche Lösungswege entstehen, erschließt sich weder anhand der sich verändernden Netzstruktur noch anhand der eingesetzten Lernalgorithmen. Nur die Ausgaben der letzten Schicht sind als Ergebnisse außerhalb des Netzes sichtbar. Bisher werden auch keine Gütekriterien ermittelt und ausgegeben, anhand derer abgeschätzt werden kann, wie aussagekräftig ein ermitteltes Ergebnis ist (interne Prüfung; Kap. 2.3.3). In Folge werden KNN teilweise auch als Blackbox bezeichnet. An Verfahren, mit denen die Entscheidungsfindung und die Ergebnisse besser nachvollzogen werden können (teils auch als erklärbare KI bzw. Explainable Artifical Intelligence – XAI bezeichnet) und an Prüfansätzen zur Bewertung der Robustheit der Verfahren wird derzeit gearbeitet (Bitkom 2019; Samek et al. 2019). Da die ermittelten strukturellen Muster nicht explizit dargestellt werden und auch Fachkräfte daraus keine Erkenntnisse ableiten können, wird das Training von KNN zwar immer als maschinelles Lernen und komplexes datenanalytisches Verfahren, aber mitunter nicht als Data-Mining im Sinne des Knowledge Discovery in Databases aufgefasst. Das ist jedoch eher eine unter Analyst/innen geführte Diskussion zur Abgrenzung unterschiedlicher Verfahren.

2.3.3 Ergebnisprüfungen

Es gibt unterschiedliche Möglichkeiten, die Richtigkeit von Data-Mining-Ergebnissen abzuschätzen und zu bewerten: zum einen verfahrensinterne Prüfansätze (vor allem klassische statistische Verfahren liefern unterschiedliche Gü-

tekennziffern), die unmittelbar an den In- und Outputkomponenten (Analysedaten und daraus abgeleitete strukturelle Muster) anknüpfen. Zum anderen gibt es verfahrensexterne Prüfansätze. Sie nehmen die Anwendung ermittelter Regeln und Modelle zur Bewertung neuer Sachverhalte in den Blick. Auch die Suche nach möglicherweise vorliegenden Fehlern erfordert externe Prüfungen. Da diese vielfältigen Ursachen haben können (unpassende Trainingsdatensätze oder Analyseverfahren, weitere zufällige Faktoren), sind diese methodisch inhaltlichen Prüfungen besonders aufwendig (Zweig 2019a, S. 150).

Interne Prüfung: Möglichkeiten und Grenzen

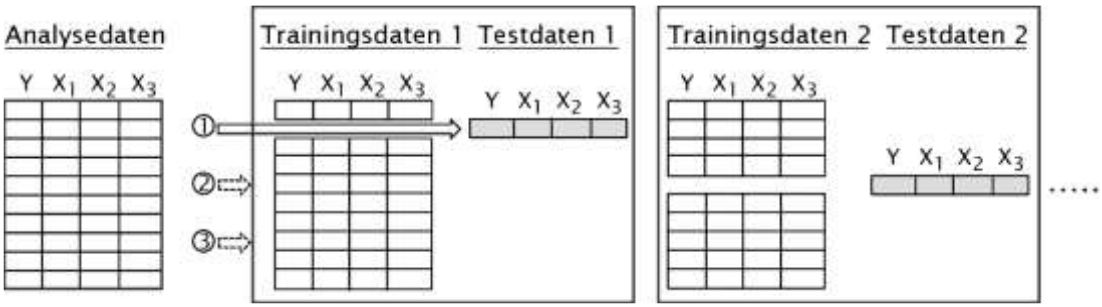
Algorithmen, die bei klassischen mathematisch-statistischen Verfahren eingesetzt werden, berechnen nicht nur definierte Parameter von Zielfunktionen und -modellen (u. a. Gruppenmittelpunkte ähnlicher Objekte, Korrelationen zwischen Merkmalen oder mögliche Minima von Zielfunktionen). Sie liefern regelmäßig auch verfahrensspezifische Kriterien (z. B. Bestimmtheitsmaße, Homogenitätskoeffizienten, Signifikanzniveaus), anhand derer abgeschätzt werden kann, wie gut das jeweils ermittelte strukturelle Muster die Analysedaten repräsentiert. Diese Kriterien steuern einerseits einzelne Prozesse der zunehmend automatisiert ablaufenden Analyseverfahren (z. B. als Lenkungs- oder Abbruchkriterien iterativer Prozesse). Andererseits können sie als Indikatoren für die Verlässlichkeit und Genauigkeit der ermittelten mathematisch-statistischen Modelle interpretiert werden und sind folglich bei der Ergebnisbewertung höchst relevant. Oft geben sie den Ausschlag, ob ein ermitteltes Ergebnis überhaupt als verwendbar eingestuft wird. Verfahrensspezifische Gütekriterien haben methodische Grenzen. In bestimmten Situationen, z. B. bei sehr selten auftretenden Sachverhalten (Anwendungsbeispiel Kap. 5.5.3) oder wenn viele Merkmale in eine Analyse einfließen, die für eine spezifische Aufgabenstellung kaum relevant sind (Problem der Überanpassung), sind einzelne Kriterien keine guten Indikatoren zur Bewertung der Modellqualität.[17]

Eine andere Prüfmöglichkeit basiert auf dem Prinzip der Kreuzvalidierung (Abb. 2.5). Dafür wird ein für eine Analyse verfügbarer Datenbestand geteilt in einen Trainings- und einen Testdatensatz. Mit den Trainingsdaten wird ein mathematisch-statistisches Modell spezifiziert, das dann die ausgeschlossenen

17 Bei klassischen statistischen Verfahren (die jeweiligen erklärenden Merkmale/Variablen werden vorab festgelegt) kann es zur Überanpassung (Overfitting) kommen, wenn für die Modellbildung zu viele Merkmale als erklärende Variablen definiert wurden, darunter auch solche, die für eine spezifische Fragestellung nicht relevant sind. Durch die Ausweitung der Anzahl erklärender Merkmale sinken die Werte der Gütekriterien nicht ab, sodass die tatsächlich geringer werdende Anpassungsgüte des Modells verschleiert wird. Dennoch fehlt dem Modell die Generalität (ermittelte Strukturen können nicht auf neue Situationen übertragen und/oder in einer größeren Grundgesamtheit sinnvoll eingesetzt werden). Das Gegenteil ist Unteranpassung (Underfitting) eines Modells, wenn tatsächlich relevante Merkmale bei einer Analyse außer Acht gelassen wurden. Auch dann fehlt dem Modell die Generalität, was die Gütekriterien jedoch anzeigen.

Testdatensätze bewertet oder prognostiziert. Durch wiederholte Neuaufsplittung der Analysedaten lassen sich mittlere Fehlerraten für das jeweilige Modell berechnen. Mit diesem Verfahren kann vor allem die Qualität von Modellen bewertet werden, die mit klassischen statistischen Verfahren spezifiziert wurden. Ergebnisbewertungen mit Kreuzvalidierungen werden mit größeren Analysedatenbeständen aufwendiger.

Abb. 2.5 Teilung eines Analysedatenbestandes zur Kreuzvalidierung



Eigene Darstellung

Auch wenn das Training von KNN derzeit mit großen Hoffnungen verbunden wird, sind Ergebnisprüfungen oftmals eine Herausforderung. Bisher liefern die jeweiligen maschinellen Lernverfahren keine internen Gütekriterien als Indikatoren für die Genauigkeit der Verfahren bzw. für die Richtigkeit der Ergebnisse. Grundsätzlich können beim KNN-Training mit überwachten Lernverfahren Fehler in der Trainingsphase verfahrensintern erkannt (wenn berechnete Zielwerte nicht mit den im Trainingsdatensatz mitgelieferten Zielwerten übereinstimmen) und durch iterative Anpassung der Stärke der Neuronenverbindungen minimiert werden (bis errechnete und mitgelieferte Zielwerte passen). Unüberwachte Lernverfahren haben diese Anpassungsmöglichkeit nicht. Auch die Berechnung von mittleren Fehlerraten durch wiederholte Neuaufsplittungen von Analysedaten in Trainings- und Testdaten ist nicht praktikabel, da KNN kontinuierlich lernen und der Lernprozess nicht rückgängig gemacht werden kann. Aufgrund der begrenzten verfahrensinternen Prüfmöglichkeiten steigt die Relevanz externer Prüfverfahren.

Externe Prüfungen: Möglichkeiten und Grenzen

Je komplexer Untersuchungsaufgaben und je weniger verfahrensinterne Prüfmöglichkeiten existieren, desto relevanter werden verfahrensexterne Prüfungen – sowohl bei symbolischen als auch bei subsymbolischen Verfahren. Externe Prüfungen können an den zugrundeliegenden Analysedaten, den einge-

setzten mustererkennenden Verfahren sowie den abgeleiteten Regeln und Modellen und den möglicherweise damit ermittelten Informationen ansetzen. Derartige Prüfungen sind in der Regel aufwendig und erfordern spezifische Fachkenntnisse. Eine Herangehensweise knüpft an die Ermittlung von Gütekriterien und Fehlerraten anhand von Testdatensätzen an. Ein anderer Ansatz ist die fachlich inhaltliche Auseinandersetzung mit den eingesetzten Analysedaten, -verfahren und den ermittelten Ergebnissen. Oftmals ergänzen sich diese Ansätze, da Gütekriterien mögliche Fehler indizieren und methodisch inhaltliche Prüfungen nach Fehlerursachen suchen.

Statistische Gütekriterien

Anhand von Testdatensätzen, die nicht bereits in ein Data-Mining-Verfahren zum Finden struktureller Muster und zur Modellparametrisierung eingesetzt wurden, kann geprüft werden, in welchem Umfang ermittelte Regeln und Modelle neue Situationen richtig erkennen und definierte Aufgaben korrekt lösen können. Vor allem bei Klassifikationsaufgaben sind folgende Gütekriterien zur externen Prüfung und Qualitätsbewertung von diversen diagnostischen und prognostischen Verfahren relevant – und in der Medizin seit Jahren genutzt (Tab. 2.1; Kap. 4.2):

- Sensitivität (Richtig-positiv-Rate): Tatsächlich Kranke werden durch ein datenbasiertes Verfahren bzw. statistisches Modell richtig als krank klassifiziert;
- Spezifität (Richtig-negativ-Rate): Gesunde werden richtig als gesund klassifiziert;
- positiver Vorhersagewert (Falsch-positiv-Rate): Gesunde werden als krank befundet;
- negativer Vorhersagewert (Falsch-negativ-Rate): Kranke werden als gesund befundet;
- Genauigkeit (Treffsicherheit): Anteil aller richtig befundeten Personen;
- Fehlerrate (Gegenstück zur Genauigkeit): Anteil aller falsch befundeten Personen.

Diese Gütekriterien ermöglichen unterschiedliche Prüfungen: Einerseits kann die Richtigkeit/Güte einzelner Entscheidungsregeln und Modelle im Zeitverlauf geprüft werden, andererseits können unterschiedliche mathematisch-statistische Verfahren (symbolische und subsymbolische Verfahren) auch mit menschlichen Fähigkeiten verglichen werden. Zur Qualitätsbewertung trainierter KNN sind diese externen Prüfungen hochrelevant, da weder verfahrensinterne Qualitätskriterien ausgegeben noch Entscheidungsprozesse transparent dargestellt werden (Blackboxproblematik). Auch können bei trainierten KNN Überanpas-

sungen auftreten (Finlayson et al. 2019; Heaven 2019), was anhand der Fehlerrate festgestellt werden kann.[18]

Tab. 2.1 Statistische Gütekriterien von Klassifikationsverfahren

		Prüfobjekte/neuer Fall		total	Gütekriterien
		krank (positiv)	gesund (negativ)		
Testergebnis	krank (positiv)	richtig positiv RP	falsch positiv FP (Fehler 2. Art)	RP+FP	*positiver Vorhersagewert* RP/(RP-FP)
	gesund (negativ)	falsch negativ FN (Fehler 1. Art)	richtig negativ RN	FN+RN	*negativer Vorhersagewert* FN/(FN+RN)
total		RP+FN	FP+RN	RP+FP+FN+RN	
Gütekriterien		*Sensitivität* RP/(RP+FN)	*Spezifität* RN/(FP+RN)	*Genauigkeit* (RP+RN)/ (RP+FP+FN+RN)	*Fehlerrate* (FP+FN)/ (RP+FP+FN+RN)

Quelle: nach https://en.wikipedia.org/wiki/Evaluation_of_binary_classifiers (13.12.2021)

Seit einigen Jahren werden für die Objekterkennung auf Abbildungen spezielle Wettbewerbe organisiert, bei denen anhand einheitlicher Testdatensätze diese Gütekriterien für unterschiedliche Klassifikationsverfahren ermittelt und verglichen werden. Laut Jones (2014) nehmen seit 2012 trainierte KNN an diesen Wettbewerben teil. Sie waren von Anfang an besser als symbolische Verfahren (Fehlerraten: subsymbolische Verfahren ca. 15 %; symbolische Verfahren ca.

18 Überanpassung kann auch bei trainierten KNN in der längerfristigen Anwendung auftreten. Nachdem ein KNN anhand eines speziell zusammengestellten Trainingsdatensatzes beispielsweise überwacht lernte, die Ziffern von 0 bis 9 auf Bildern zu erkennen, kann es im Anschluss neue handschriftliche Ziffern zuordnen. Die Erkennungsleistung des KNN steigt in der Trainingsphase und darüber hinaus zunächst kontinuierlich (Fehlerrate sinkt) bis sie ein gewisses Niveau erreicht (Fehlerrate erreicht ein Minimum, auch als Sättigungsphase bezeichnet). Mitunter steigt die Fehlerrate danach wieder, wenn sich das KNN zu sehr an die kontinuierlich vorgelegten Anwendungsfälle anpasst und sich nicht mehr an den ursprünglich vorgelegten Formen orientiert. Mitunter reichen bereits geringfügige Veränderungen aus. Um eine mögliche Überanpassung zu erkennen, sollte in regelmäßigen Abständen mit einem neuen Testdatensatz die Fehlerrate ermittelt werden. Steigt sie im Zeitverlauf, ist dies ein Hinweis auf Überanpassung. Dann sollte ein neues KNN trainiert und eingesetzt werden.

25 %) und erreichen bei einfachen Objekten seit 2014 in etwa menschliches Niveau. Dieserart Wettbewerbe sind nicht auf die Erkennung fachspezifischer komplexer Objekte ausgerichtet (z. B. die Befundung von Mammografieaufnahmen).

Liu et al. (2019) haben im Rahmen einer umfangreichen Metaanalyse erstmals systematisch Publikationen zu Studien erfasst, die unterschiedliche datenbasierte Verfahren sowie menschliche Fachexpertisen zur Erkennung von Auffälligkeiten auf medizinischen Abbildungen thematisierten. Sie fanden nur wenige methodisch hochwertige Studien (insgesamt 14), die statistische Gütekriterien vergleichbar auswiesen und dadurch extern validierbar waren. Vor allem Studien, in denen trainierte KNN eingesetzt wurden, waren häufig schlecht dokumentiert. Mitunter wurden nur einzelne statistische Gütekriterien ausgewiesen, teilweise ohne diese in einen analytischen Kontext zu stellen und ohne auf deren Grenzen der Aussagefähigkeit hinzuweisen oder die Ergebnisse mit anderen Verfahren zu vergleichen.[19] Die Ergebnisse der wenigen methodisch hochwertigen Studien deuten darauf hin, dass in unterschiedlichen Krankheitskontexten die anhand von statistischen Gütekriterien definierte diagnostische Leistung (vor allem Sensitivität, Spezifität und Fehlerrate) von trainierten KNN mit der von Radiologinnen und Radiologen vergleichbar sind (Ching et al. 2018, S. 33; Liu et al. 2019, S. 21).

Um methodisch korrekter vorgehen zu können, werden seit einigen Jahren für definierte Untersuchungsaufgaben Trainings- und Testdatensätze zusammengestellt.[20] Mit den Testdatensätzen können nach der Trainingsphase auch Gütekriterien für unterschiedliche Verfahren unter gleichen Bedingungen retrospektiv ermittelt und verglichen werden. Bei klassischen symbolischen Verfahren sollten sich die jeweiligen statistischen Kennziffern über die Zeit nicht verändern, da die Zuordnung immer nach der gleichen Berechnungsvorschrift erfolgt. Da sich bei subsymbolischen Verfahren diese Kennziffern im Zeitverlauf ändern können, müssen Methoden zur Qualitätsbewertung und zum Leistungsvergleich entsprechend weiterentwickelt werden. Liu et al. (2019) fordern entsprechend neue Berichterstattungsstandards für externe Ergebnisvalidierungen, die insbesondere die spezifischen Herausforderungen von trainierten KNN berücksichtigen. Auch sollten Studiendesigns methodisch weiterentwickelt und die Studienqualität verbessert werden, um mehr Vertrauen zu dieser vielversprechenden Technologie aufzubauen und Ergebnisse extern bewerten zu können.

19 Insbesondere bei (sehr) selten auftretenden Ereignissen werden Vorhersagewerte oftmals intuitiv fehlinterpretiert. Im Rahmen der vom Max-Planck-Institut für Bildungsforschung herausgegebenen Unstatistik des Monats werden regelmäßig Fehlinterpretationen diskutiert (www.mpib-berlin.mpg.de/de/presse/dossiers/unstatistik-des-monats; 13.12.2021).

20 Umfangreiche themenspezifische Listen von Trainings- und Testdatensätzen gibt es beispielsweise unter wikipedia.org/wiki/List_of_datasets_for_machine-learning_research (13.12.2021)

Methodisch inhaltliche Prüfungen und Data-Mining-Grenzen

Als Ursachen für Fehler im Rahmen von Data-Mining-Prozessen kommen die jeweils eingesetzten Trainingsdaten, die verwendeten Analyseverfahren und diverse weitere zufällige Faktoren in Frage (Zweig 2019a, S. 150). Einige Aspekte sollen nachfolgend angerissen werden, um die Bedeutung von Fachkenntnissen und methodisch inhaltlichen Fehlerprüfungen zu unterstreichen. Derartige Fehlerprüfungen erscheinen umso dringender, je einfacher vielfältige Datenbestände mit voreingestellten Analysetools von datenanalytischer Software verarbeitet werden können und die im Hintergrund arbeitenden Algorithmen so programmiert sind, dass die immer Ergebnisse liefern (Kasten 2.3).

Kasten 2.3 Beispielhafte externe Prüfung voreingestellter Analysetools

Seit ca. 25 Jahren wird in der neurowissenschaftlichen Forschung die funktionale Magnetresonanztherapie (fMRT) eingesetzt, ein bildgebendes Verfahren zur Messung der Hirnaktivität. Methodisch anerkannt ist die Zusammenführung von fMRT-Messdaten mehrerer aktiver Probanden und der Vergleich mit ebenfalls zusammengeführten fMRT-Daten mehrerer ruhender Probanden. Abweichungen in den Aktivitätsmustern von aktiven und ruhenden Probandengruppen zeigen an, welche Hirnregion bei der jeweils ausgeführten Aktivität wie stark beteiligt ist. Dieses Vorgehen wird inzwischen mittels in fMRT-Analysesoftware integrierten Algorithmen weitgehend automatisiert realisiert. In sich auf fMRT-Datenanalysen stützenden wissenschaftlichen Publikationen der letzten Jahre dominierten drei Analyseprogramme. Alle nutzten für den Vergleich einen synthetischen Datensatz einer ruhenden Probandengruppe. Die Programme sind anerkannt und werden bei Forschungsprojekten regelmäßig eingesetzt, ohne dass bei jedem Einsatz jedes methodische Detail neu geprüft wird. Auch Peer-Reviews entsprechender Publikationen forderten keine explizite Validierung der Vergleichsdatensätze mehr ein.

2016 prüften Analysten diese Vorgehensweise (Eklund et al. 2016). Sie verglichen die voreingestellt hinzugezogenen synthetischen Datensätze mit inzwischen vielfach öffentlich verfügbaren realen fMRT-Daten ruhender Probanden und stießen auf erhebliche Abweichungen. Sie stellten fest, dass in einzelnen Studien kaum noch geprüft wurde, inwiefern der vom Algorithmus standardmäßig genutzte Datensatz alle Voraussetzungen für einen korrekten Vergleich mit den jeweiligen Studiendaten erfüllt. Da diese Voraussetzungen oft nicht erfüllt waren, die Algorithmen aber so designt sind, dass sie trotzdem eingesetzt werden können und Ergebnisse liefern, produzierten sie häufiger falsch positive Werte, d. h., sie wiesen Hirnaktivität aus, wo ei-

gentlich keine war. Die Überprüfung deutete auf deutliche Überschreitungender normalerweise tolerierten Fehlerraten hin. Dies war kein Fehler des Algorithmus an sich, vielmehr wurden voreingestellte Vergleichsdatensätze für die Berechnung herangezogen, die situativ nicht passten. Die Autoren wiesen nachdrücklich darauf hin, Algorithmen nicht blind anzuwenden, sondern sich vor jedem Einsatz mit deren spezifischer Methodik und deren jeweiligen Grenzen auseinanderzusetzen.

Bei der methodischen Prüfung ist es wichtig, sich der Grenzen datenbasierten Vorgehens bewusst zu sein: Mit Data-Mining-Verfahren können nur solche Muster und Strukturen gefunden werden, die in den jeweiligen Analysedatensätzen enthalten sind. Neue Phänomene sind in alten Datenbeständen oftmals nicht, nicht gut oder genau genug enthalten (z. B. dürften Datensätze von menschlichen Blutproben, die vor 2020 erhoben wurden, keine Informationen zu SARS-CoV-2-Erreger enthalten). Hochkomplexe Sachverhalte sind oft nicht in ausreichender Detailgenauigkeit enthalten und Modelle als vereinfachte Darstellungen können nur Teilaspekte abbilden (z. B. können Wettermodelle lokale Entwicklungen in ihrer Dynamik nur schwer abbilden). Aufgrund von vereinfachten Darstellungen können extrahierte Strukturen zu Fehlinterpretationen führen, weil zufällige oder indirekte Beziehungen abgebildet werden und Korrelationen als Kausalitäten gedeutet werden (der Klassiker ist die Korrelation zwischen menschlichen Geburtenraten und der Anzahl an Storchenpaaren in unterschiedlichen europäischen Regionen [Matthews 2000]). Sehr seltene Ereignisse sind in mehrfacher Hinsicht eine datenanalytische Herausforderung: Da Trainingsdaten meist nur Stichproben sind, besteht die Gefahr, dass seltene Ereignisse unterrepräsentiert oder gar nicht enthalten sind (das Fehlen solcher Ereignisse, heißt nicht, dass es diese nicht gibt) oder Gütekriterien fehlinterpretiert werden.

Trainierte KNN sind für methodisch-inhaltliche Prüfungen eine besondere Herausforderung. Wegen der Blackboxdarstellungen kann das Vorgehen nicht direkt anhand von Parametern, Formeln oder Kennziffern nachvollzogen werden. Verfahrensspezifische Fehler, wie z. B. Fehlanpassungen in der Trainingsphase, können nur schwer gefunden werden.

In der methodischen Diskussion zu den Stärken und Grenzen von Klassifikationsverfahren wird mitunter die Fehleranfälligkeit als Bewertungskriterium herangezogen. Eher allgemein logisch wird argumentiert, dass im Gegensatz zum Menschen Computer und Algorithmen nie müde und unaufmerksam werden und in Folge die Fehlerquote nicht müdigkeitsbedingt steigen würde. Aufgrund der hohen verfügbaren Rechenleistungen könnten vor allem KNN der neuesten Generation in Sekundenbruchteilen bereits kleinste Veränderungen erfassen, die selbst geschulten Experten mitunter leicht entgehen würden. Diesem Argument wird teilweise entgegengehalten, dass insbesondere trainierte KNN

bereits durch geringfügige Datenveränderungen auch leicht zu täuschen und anfällig für Überanpassungen seien. Einige Analyst/innen bewerten diesen Ansatz deshalb derzeit dennoch als fehleranfälliger (auch in Bezug auf missbräuchliche Manipulationen z. B. durch Hackerangriffe) und weniger robust im Vergleich zu Fachkräftebewertungen (Finlayson et al. 2019; Heaven 2019). Diverse Verfahrensvergleiche zeigen, dass bei komplexen Sachverhalten in der Regel weder mathematisch-statistische Verfahren noch die Expertise von Fachkräften vollständige Fehlerfreiheit garantieren können. Je nach Anwendungsbereich ist deshalb die Frage nach den Folgendimensionen und der Minimierung möglicher Schäden beim Einsatz derartiger Verfahren von besonderer Bedeutung. Dafür wird dann oftmals differenziert, ob Fehler 1. oder 2. Art (Tab. 2.1) ggf. tolerierbar seien. Die Diskussion zu den Stärken und Grenzen unterschiedlicher analytischer Ansätze ist nicht fundamental neu, sie wird in der Medizin kontinuierlich geführt (Kap. 4).

Eine weitere derzeit im Kontext von komplexen Datenanalysen oder Big Data thematisierte mögliche Folgedimension ist die Diskriminierung von Einzelpersonen durch den Einsatz komplexer Analyseverfahren. Sie ist einerseits gesetzlich verboten (Allgemeines Gleichbehandlungsgesetz – AGG). Andererseits lässt sie sich bei komplexen Datenanalysen kaum per se ausschließen. Wenn in einer Gesellschaft Formen von Diskriminierung real existieren und Analysedatensätze diese Realität korrekt abbilden, besteht die Gefahr, dass diese Diskriminierung durch die ermittelten Regeln und Modelle reproduziert werden. Um möglicherweise vorliegende Diskriminierung zu be- oder widerlegen, sind situative methodisch inhaltliche Prüfungen erforderlich (ausführlich z. B. Orwat 2019; TAB 2020).

Aufgrund der zunehmenden Möglichkeiten datenbasierte Muster zu erkennen, Regeln und Modelle abzuleiten und mittels Algorithmen in neuen Situationen bei vielfältigen Entscheidungsprozessen einzusetzen und aufgrund der möglichen Folgen sowie der begrenzten Möglichkeiten der Prüfung für letztendlich betroffene Personen wird seit einigen Jahren die Forderung nach einer stärkeren Überwachung und einer kompetenten Prüfinstanz für algorithmische Systeme lauter (DEK 2019, S. 173 ff.), mitunter auch als Algorithmen-TÜV bezeichnet (Zweig 2019b, S. 8 ff.). In der Medizin werden entsprechende Verfahren seit Jahren auf- und ausgebaut (Kap. 4.2).

2.3.4 Weiterverwendung von Data-Mining-Ergebnissen

Die Suche nach Zusammenhängen, Ähnlichkeiten, Unterschieden und Besonderheiten in Datenbeständen und die Auseinandersetzung mit der Generalisierbarkeit der ermittelten Zusammenhänge und Strukturen sind seit jeher Formen wissenschaftlichen Vorgehens, jedoch bei weitem nicht auf wissenschaftliche

Kontexte begrenzt. Entsprechende Vorgehensweisen sind bei vielfältigen Entscheidungsprozessen von großer Bedeutung. Die Digitalisierung hat diesbezüglich eine erhebliche Schubkraft, weil immer umfangreichere Datenbestände generiert werden und weil maschinenlesbare Formate von Menschen schwer unmittelbar erfasst werden können. Maschinelle bzw. zunehmend automatisierte Entscheidungssysteme (Automated Decision Making – ADM) gewinnen an Relevanz. Auch wenn es erhebliche Schnittmengen beim datenbasierten Vorgehen in wissenschaftlichen und operativen Bereichen gibt, sind die Anwendungskontexte und Zielsetzungen keinesfalls deckungsgleich.

In wissenschaftlichen Bereichen werden seit jeher Ausschnitte aus der realen Welt vermessen und Daten analysiert, um Strukturen und Muster zu erkennen, Erkenntnisse zu gewinnen, oder Regeln bis hin zu (Natur-)Gesetzen abzuleiten. Wissenschaftliches Arbeiten bedeutet, bisher Unbekanntes zu erschließen und im Wortsinn Wissen zu schaffen, Hypothesen zu prüfen, deren Richtigkeit zu be- oder zu widerlegen (auch mit dem Begriff der Evidenz assoziiert). Eine größer werdende Datenbasis kann beim wissenschaftlichen Vorgehen die empirische Evidenz bezüglich bestimmter Zusammenhänge (Korrelationen) zwischen Sachverhalten erhöhen. Beweise für die Richtigkeit dieses Zusammenhangs sollten möglichst auf andere Art und Weise erbracht werden. Derartige Beweisführungen erfordern regelmäßig weitere wissenschaftliche Aktivitäten, die weit über Data-Mining-Prozesse hinaus gehen. Mitunter wird diese wissenschaftliche Auseinandersetzung auch durch den Übergang des »know how« zum »know why« (Zeleny 1987) bzw. des Schrittes vom Finden von Korrelationen zur Suche nach Kausalitäten umschrieben. Im historischen Beispiel (Kap. 2.1) half John Snow die Lokalisierung des Zentrums des Choleraausbruchs zum Ersten bei der Entscheidung, den im Zentrum stehenden Wasserbrunnen sofort stillzulegen, auch wenn er seine Hypothese zur Krankheitsursache nicht zweifelsfrei beweisen konnte (*Entscheidungshilfe*). Zum Zweiten führte die Lokalisierung des Zentrums der Epidemiologie zu weiteren Untersuchungen bezüglich der Krankheitsursachen, die seine Erregerhypothese fundieren sollten (*Wissenserweiterung*). Zum Dritten wurden räumliche Analysen als Entscheidungshilfen verstetigt – John Snow gilt als einer der diesbezüglichen Pioniere. Dazu werden die einzelnen analytischen Schritte in Algorithmen übersetzt und inzwischen in Software eingebunden (*Operationalisierung datenanalytischer Verfahren*). Wenn das gelingt, können einzelne datenanalytische Verfahren in neuen Situationen der gleichen Untersuchungsaufgabe eingesetzt werden und zumindest situative Informationen als Entscheidungshilfen liefern. Die Automatisierung von Entscheidungsprozessen ist dann ein nächster Schritt. Je nach Betrachtungsweise geht auch die Entwicklung algorithmischer Softwaresysteme zur Entscheidungsfindung über den eigentlichen Data-Mining-Prozess hinaus.

Rechtlich haben zum einen Data-Mining-Aktivitäten zu wissenschaftlichen Forschungszwecken und zum anderen automatisierte Entscheidungssysteme gewisse Sonderstellungen. Die derzeitige rechtliche Situation und die gesellschaftlichen Herausforderungen in Bezug auf die Verbesserung des Zugangs zu existierenden Datenbeständen für komplexe Datenanalysen sowie die Regulierung des Umgangs mit Data-Mining-Ergebnissen werden im nächsten Kapitel 3 thematisiert.

3 Rechtliche und normative Aspekte

Seit einigen Jahren wird der Begriff Data-Mining auch aus rechtlicher Sicht diskutiert und im Allgemeinen mit komplexen Datenanalysen assoziiert (u. a. Nohr 2017; Triaille et al. 2014). Seit 2019 findet er über das Urheberrecht Eingang in das nationale und europäische Rechtssystem. Dort wird Data-Mining definiert als »eine Technik für die automatisierte Analyse von Daten in digitaler Form, mit deren Hilfe Informationen unter anderem – aber nicht ausschließlich – über Muster, Trends und Korrelationen gewonnen werden können« (Art. 2 Abs. 2 Richtlinie 2019/790/EU[21]). Diese Formulierung entspricht der Auffassung von Data-Mining im engeren Sinn innerhalb dieses Berichtes. Betrachtet man Data-Mining als Prozess im weiteren Sinn, wird deutlich, dass die Erhebung, Bereitstellung und Analyse von Daten sowie der Umgang mit den resultierenden Informationen, Regeln, Modellen oder algorithmischen Systemen weit über den Regulierungshorizont des Urheberrechts (Kap. 8.1) hinaus geht. Einige Aspekte wie der Datenschutz oder die Datennutzung in öffentlichen Aufgabenbereichen werden bereits seit Jahren reguliert, wobei Data-Mining meist unter dem allgemeinen Oberbegriff Datenverarbeitung subsumiert wird. Der Umgang mit daraus hervorgehenden Informationen, datenanalytischen Werkzeugen und digitalen Anwendungen nimmt erst schrittweise Gestalt an.

Die Erhebung und Analyse von Daten durchdringen das tägliche Leben und das Miteinander zunehmend, betreffen sowohl einzelne Personen als auch Personengruppen oder die Gemeinschaft, öffentlichen Aufgabenbereiche sowie gewerbliche Aktivitäten. Nach einem Einblick in datenbezogene rechtliche Grundstrukturen (Kap. 3.1) wird erst der Bereich der Daten ohne Personenbezug betrachtet. In diesem können die Rechte und Möglichkeiten datenverarbeitender Stellen ins Zentrum der Betrachtung gestellt werden, ohne dass die Rechte von Personen als von Datenverarbeitung betroffene berücksichtigt werden müssen. *Datenzugangsstrukturen* werden beispielhaft anhand der nationalen Geodateninfrastruktur öffentlicher Einrichtungen mit dem derzeitigen Open-Data-Konzept umrissen (Kap. 3.2). Dieser Datenzugang ist oft eine notwendige Bedingung für Data-Mining-Aktivitäten. Der Bereich der personenbezogenen bzw. -beziehbaren Daten hat eine größere rechtliche Komplexität, denn natürliche Personen haben Grundrechte, die bei jeglicher Datenverarbeitung zu achten und zu schützen sind. Das Datenschutzrecht hat den Anspruch, Datenanalytik unter Achtung der Grundrechte betroffener Personen zu ermöglichen. Unterschiedliche Schutzmaßnahmen sowie grundrechteschützende Datenzugangs-

21 Richtlinie (EU) 2019/790 über das Urheberrecht und die verwandten Schutzrechte im digitalen Binnenmarkt und zur Änderung der Richtlinien 96/6/EG und 2001/29/EG

strukturen und Analyseansätze (Privacy Preserving Data Mining) werden skizziert. Diese Maßnahmen und Ansätze bilden das Fundament, auf dem Data-Mining-Aktivitäten im Rahmen der wissenschaftlichen Forschung Sonderkonditionen im Datenschutzrecht gewährt werden (Kap. 3.3). Der Umgang mit den aus Data-Mining-Prozessen ggf. hervorgehenden Informationen, datenbasierten Werkzeugen, Produkten und Services ist ein Gegenstand der derzeitigen regulativen Debatten (Kap. 3.4).

3.1 Datenbezogene Grundstrukturen

Daten versus Information

Der Umgang mit Daten wird durch vielfältige Rechtsnormen angesprochen und reguliert. Dennoch gibt es bisher keine Definition, die für alle Rechtsbereiche gleichermaßen gilt. In einigen Rechtsnormen werden die Begriffe Daten und Informationen weitgehend synonym verwendet (z. B. DSGVO oder Umweltinformationsgesetze auf Bundes- und Länderebene), in anderen wird der Datenbegriff stärker mit strukturierten, zumindest temporär gespeicherten, maschinenlesbaren Aufzeichnungen verknüpft (z. B. § 202a StGB[22]) und der Informationsbegriff eher mit für Menschen direkt verständlichen Aussagen verbunden (Riehm 2018, S. 74). Laut den für öffentliche Einrichtungen auf Bundes- und Landesebene gültigen Informationsfreiheitsgesetzen[23] sind Informationen jegliche, amtlichen Zwecken dienende Aufzeichnungen (Schriften, Tabellen, Diagramme, Bilder, Pläne, Karten, Tonaufzeichnungen, nicht aber Entwürfe und Notizen), unabhängig von der Art der Speicherung. Auch bei der Definition von Metadaten wird die begriffliche Nähe von Daten und Informationen deutlich. Laut Geodatenzugangsgesetzen[24] auf Bundes- und Landesebene sind Metadaten Informationen, die (Geo-)Daten beschreiben und es möglichen, (Geo-)Daten und (Geo-)Datendienste zu ermitteln, in Verzeichnisse auszunehmen und zu nutzen (§ 3 Abs. 2 GeoZG).

Aus der Schutz- bzw. Geheimhaltungsperspektive werden die Begriffe Daten und Informationen bereits mit sehr kleinen Elementen oder Einheiten in Verbindung gebracht. Bereits Einzelangaben sind als Privat- oder Geschäftsgeheimnisse durch das Strafgesetzbuch geschützt (§ 203 StGB). Auch nach Art. 4

22 Strafgesetzbuch (StGB)
23 Gesetz zur Regelung des Zugangs zu Informationen des Bundes (Informationsfreiheitsgesetz – IFG) S. 2722) (BGBl. I S. 1328)
24 Bundesebene: Gesetz über den Zugang zu digitalen Geodaten (Geodatenzugangsgesetz – GeoZG) (BGBl. I S. 278); Übersicht und Zugang zu den Geodatenzugangsgesetzen auf Landesebene z. B. unter www.bmu.de/themen/bildung-beteiligung/umweltinformation/umweltinformationsgesetz/uebersicht-der-geodatenzugangsgesetze-der-bundeslaender/ (13.12.2021)

Abs. 1 DSGVO werden alle Daten bzw. Informationen, die auf natürliche Personen bezogen oder beziehbar sind, unabhängig von der Größe eines Datensatzes erfasst. Zudem wird die Leistung zur Erstellung komplexer Datensätze und -bestände zum Teil durch das Urheberrecht geschützt (Kap. 8.1). Dennoch ist Data-Mining zu wissenschaftlichen Forschungszwecken möglich, sofern diese Daten dafür zugänglich sind.

Der Rechtsraum zum Umgang mit Daten kann in erheblichem Maße anhand von zwei Spezifika erschlossen werden: zum einen über die Bezüge zu vielfältigen Sachverhalten einschließlich Personen (technisch allgemein als *Datenobjekte* aufgefasst), zum anderen über die Tatsache, dass kaum diese Datenobjekte selbst, sondern vor allem andere Akteure die Möglichkeit haben, Daten zunehmend detailreich erheben, dauerhaft in großen Mengen halten sowie über diverse Bezüge vielfältig verknüpfen, verändern und ergänzen sowie u. a. Data-Mining durchführen zu können (rechtlich als *datenverarbeitende Stellen* aufgefasst).

Datenobjekte und Datenbereiche

Datenobjekte sind höchstdivers und vielfältig, jedoch sind einzelne Personen aufgrund ihrer subjektiven (Grund-)Rechte besondere Datenobjekte. *Natürliche Personen* haben u. a. Grundrechte auf Privatheit und informationelle Selbstbestimmung bzw. Datenschutz. Ähnlich haben am Wettbewerb teilnehmende Organisationen und Unternehmen als *juristische Personen des privaten Rechts* Grundrechte auf Berufsfreiheit bzw. Geschäftsgeheimnisse. *Juristische Personen des öffentlichen Rechts* (Behörden, Ämter oder Gebietskörperschaften) haben im Rahmen ihrer definierten Aufgaben teilweise Amts-, Dienst- oder sogar Staatsgeheimnisse zu wahren. Neben ihren Geheimhaltungsrechten haben Personen in der nationalen Rechtsordnung auch Grundrechte auf Eigentum an materiellen und immateriellen Gütern (Kap. 8.1). Keines dieser (Grund-)Rechte ist schrankenlos. Jedes kann bei überwiegendem öffentlichem Interesse gesetzlich begrenzt werden. Alle Grundrechte sind entsprechend der jeweiligen normativen Grenzen bei jeglicher Form der Datenverarbeitung einschließlich Data-Mining zu beachten und zu schützen.

In Folge wird der Datenraum rechtlich in erheblichem Maße zweigeteilt: in den Bereich der Daten *ohne* Bezüge zu einzelnen Personen (z. B. Geodaten; Kap. 3.2) und in den Bereich der Daten *mit* Bezügen zu einzelnen Personen (Kap. 3.3). Auch heute bilden diese Bereiche gewisse Pole. Jedoch werden die dazwischen liegenden Graubereiche immer größer, weil eine Personenbeziehbarkeit bei sehr vielen Datensätzen möglich wird (Kap. 3.3.5) und teilweise sogar Bezüge zu unterschiedlichen Personen existieren (in besonderem Maße z. B. bei Leistungsabrechnungsdaten im nationalen Gesundheitssystem; Kap. 4.4.1). Trotz der zunehmenden Graubereiche hilft diese Strukturierung, um den

Rechtsraum grundsätzlich zu erschließen und sich der Frage zu nähern, inwiefern Data-Mining rechtlich erfasst, ermöglicht oder begrenzt wird.

Einzelne Personen sind zwar besondere Datenobjekte. Da sie in der Regel keinen Zugang zu Daten von vielen anderen Einzelpersonen haben, können sie selbst kaum datenbasierte Muster und Strukturen mittels Data-Mining extrahieren. Das können nur *datenverarbeitende Stellen*, sofern sie entsprechende Datennutzungsrechte haben.

Datenverarbeitende Stellen

Aus datenschutzrechtlicher Sicht sind datenverarbeitende Stellen verantwortliche oder im Auftrag handelnde Personen, die Daten erheben, ordnen, speichern, verändern, verwenden, verbreiten, verknüpfen, löschen oder vernichten (Art. 4 DSGVO). Aus Data-Mining-Perspektive relevant sind vor allem juristische Personen entweder des privaten oder des öffentlichen Rechts, da de facto nur diese über die notwendigen Data-Mining-Ressourcen verfügen (Nutzungsrechte und Know-how).

Juristische Personen des privaten Rechts sind gewerblich oder gemeinnützig agierende Unternehmen oder Organisationen. Sie haben Grundrechte auf Berufs- bzw. Vereinigungsfreiheit und können im Rahmen des geltenden Rechts u. a. Data-Mining-Projekte durchführen, digitale Analysewerkzeuge entwickeln, die sie im Wettbewerb mit anderen mit oder ohne Gewinnerzielungsabsicht vermarkten können. Einige Internetunternehmen bieten mit ihren digitalen Plattformen Möglichkeiten, um viele geschäftliche und private Kunden einzubinden, wodurch teils bewährte, aber auch neue digitale Geschäftsmodelle vorangetrieben werden können (teilweise auch als Plattformökonomie bezeichnet, ausführlicher z. B. Rüchardt 2019). Mit ihnen gehen unterschiedliche gesellschaftliche Herausforderungen einher, u. a., weil marktbeherrschende Stellungen Wettbewerbsstrukturen verzerren und globale Akteure sich vielfältigen nationalen Regelungen entziehen können. Gemeinnützige Organisationen setzen mit ihren digitalen Plattformen auf mehr Offenheit und gemeinsamer Nutzung von Daten (Kasten 3.3), Standards und Programmen, wodurch ebenfalls neue Geschäftsmodelle entstehen, die nicht auf geistigem Eigentum sowie gewerblichen Entwicklungs- und Vermarktungsprozessen beruhen. Für juristische Personen des privaten Rechts gelten bei der Verarbeitung personenbezogener Daten Vorgaben der DSGVO vollumfänglich, d. h., Betroffene haben auf vertraglicher Basis in die Datenverarbeitung eingewilligt und den datenverarbeitenden Stellen Nutzungsrechte gewährt (ausführlicher z. B. Riehm 2018). Datenverarbeitende Stellen des privaten Rechts haben ihrerseits ein Recht auf Geschäftsgeheimnisse, das Datenbestände und Verarbeitungsvorgänge im Rahmen ihrer Geschäftstätigkeiten (bei personenbezogenen Daten unter Achtung der Vorgaben der DSGVO) einschließt. Diese Geheimhaltungsrechte und -pflichten

werden durch unterschiedliche Informations-, Offenlegungs- und Meldepflichten begrenzt.

Juristische Personen des öffentlichen Rechts sind u. a. Behörden, Ämter [Gebiets-]Körperschaften, die mit öffentlichen Aufgaben betraut sind (nachfolgend vereinfachend als öffentliche Einrichtungen bezeichnet). Diese öffentlichen Aufgaben und die dafür ggf. nötige Datenerhebung und -verarbeitung werden gesetzlich definiert und wesentlich aus öffentlichen Mitteln (einschließlich solidarisch finanzierter gesetzlicher Sicherungssysteme; Kapitel 5) finanziert. Teilweise dürfen/müssen die jeweiligen Einrichtungen für ihre Leistungen Gebühren erheben. In Deutschland sind einige Aufgaben auf kommunaler, andere auf Landes- oder auf Bundesebene zu erfüllen, sodass unterschiedliche Landes- oder Bundesgesetze die Vorgehensweisen spezifizieren. Öffentliche Einrichtungen haben in ihrem Aufgabenfeld regelmäßig eine Sonder- oder sogar Monopolstellung. Um an öffentliche Aufgaben angrenzende wettbewerbliche Marktstrukturen nicht zu verzerren, sollen sie nicht jenseits ihrer definierten Aufgaben agieren.[25] Datenbezogene Kooperationen mit juristischen Personen des privaten Rechts können ggf. vertraglich vereinbart werden. Öffentliche Einrichtungen haben einerseits Dienst-/Amts-/Staatsgeheimnisse zu wahren und werden andererseits zunehmend zur Offenlegung und Transparenz verpflichtet. Um deren Datenbestände rechtskonform zugänglich zu machen, werden gegenwärtig bereichsspezifische *Datenzentren und -infrastrukturen* aufgebaut.

Nicht alle datenverarbeitenden Stellen sind eindeutig entweder dem einen oder dem anderen Bereich zuzuordnen. Teilweise tragen wirtschaftlich agierende Unternehmen zur Erfüllung öffentlicher Aufgaben bei (z. B. erheben Vermessungsingenieure amtliche Geodaten oder liefern medizinische Einrichtungen Daten zur Überwachung der gesundheitlichen Situation), teilweise agieren öffentlich finanzierte Einrichtungen in Wettbewerbsstrukturen (z. B. Krankenhäuser oder Krankenkassen) oder relativ frei (z. B. wissenschaftliche Einrichtungen). Wissenschaftlichen Einrichtungen werden bei Data-Mining-Aktivitäten mit geschützten Daten Sonderrechte gewährt (Kap. 3.3.4 u. 8.1).

In der weiteren normativen Differenzierung wird teilweise zwischen vorrangig datenbereitstellenden und datenauswertenden Stellen sowie zwischen verantwortlichen und auftragsverarbeitenden Stellen unterschieden. Auch bei dieser Differenzierung gibt es Überschneidungen und Graubereiche. Dennoch

25 Beispielsweise hat der Deutsche Wetterdienst (DWD) als Bundesoberbehörde die Aufgabe, kontinuierlich umfangreiche Wetterdaten zu erheben, öffentlich zugänglich zu machen, sie zu analysieren und die Öffentlichkeit vor Unwetter zu warnen (§ 4 DWD-Gesetz). Der Bundesgerichtshof (2020) entschied, dass der DWD jenseits der kostenlosen Unwetterwarnungen, keine vollständigen Wetterinformationen kosten- und werbefrei anbieten darf, da dies nicht in seinem gesetzlichen Aufgabenbereich liege und den Wettbewerb verzerren würde.

hilft auch hier die Strukturierung, um sich den Rechtsraum grundsätzlich zu erschließen und sich der Frage zu nähern, wer in welcher Form Data-Mining befördern, realisieren und ggf. Mehrwert generieren kann.

Daten, Analysen und IT-Systeme mit besonderer Kritikalität

In vielen gesellschaftlichen Bereichen werden Datensegmente mit besonderer Relevanz, aber auch mit besonderer Kritikalität erkannt und in Folge die Datenerhebung teilweise als öffentliche oder hoheitliche Aufgaben definiert, mitunter die Datenverarbeitung spezifisch reguliert, besonders geschützt und teilweise begrenzt. Beispiele für gesetzlich definierte kritische und besonders kritische Datensegmente sind hochaufgelöste Geodaten [Kap. 3.2] oder personenbezogene Daten besonderer Kategorie [Kasten 3.6]). Auch bei einzelnen datenanalytischen Verfahren wird zunehmend eine besondere Kritikalität unterstellt, spezifische Folgenabschätzungen verlangt (Kap. 3.3.2) und für risikoreiche Verfahren eine stärkere Regulierung diskutiert (Kap. 3.4.3). Zudem werden in bestimmten gesellschaftlich relevanten Bereichen datenverarbeitende Stellen insgesamt als kritisch bewertet und deren gesamte IT-Systeme als *kritische Infrastrukturkomponenten* aufgefasst, gesichert und überwacht (Kasten 3.1).

Kasten 3.1 Kritische Infrastrukturen

Der Begriff *Kritische Infrastrukturen* (KRITIS) wird vor allem aus der Sicherheitsperspektive verwendet. Die Richtlinie 2008/114/EG[26] fasst darunter Anlagen und (Teil-)Systeme, die für die Realisierung und Aufrechterhaltung wichtiger gesellschaftlicher Aufgaben wie z. B. für die gesundheitliche Versorgung oder die Sicherung des Wohlergehens der Bevölkerung bedeutsam und folglich in besonderem Maße zu schützen sind. Da diese zunehmend auf sicher funktionierende IT-Systeme einschließlich kontinuierlicher Datenbereitstellung angewiesen sind, müssen auch diese besonders geschützt werden. In der diesbezüglichen Sicherheitsarchitektur hat das Bundesamt für Sicherheit in der Informationstechnik (BSI) eine zentrale Funktion.[27] Diese Sicherheitsarchitektur wird stufenweise definiert. In Deutschland gibt es gegenwärtig neun KRITIS-Sektoren mit insgesamt 29 Branchen, die eine wichtige Bedeutung für das staatliche Gemeinwesen haben, bei deren Beeinträchtigung nachhaltig wirkende Versorgungsengpässe, erhebliche Störungen der

26 Richtlinie 2008/114/EG über die Ermittlung und Ausweisung europäischer kritischer Infrastrukturen und die Bewertung der Notwendigkeit, ihren Schutz zu verbessern

27 Organisation und Aufgaben des BSI sind definiert im Gesetz über das Bundesamt für Sicherheit in der Informationstechnik (BSI-Gesetz – BSIG)

öffentlichen Sicherheit oder andere dramatische Folgen zu erwarten sind.[28] Der Gesundheitssektor ist seit 2017 einer davon. Krankenhäuser mit jährlich mindestens 30.000 vollstationären Behandlungsfällen gelten als KRITIS-Betreiber (DKG 2017, S. 10 f.), d. h., sie sind u. a. beim BSI registrierungs- und überwachungspflichtig, müssen bis 2019 ihre IT nach dem Stand der Technik angemessen absichern und die Sicherheit zweijährlich nachweisen (BSI 2017, S. 16). Die Deutsche Krankenhausgesellschaft und das BSI haben sich Ende 2019 auf generelle Cybersicherheitsstandards für Krankenhäuser geeinigt.[29] Ab 2022 sind alle Krankenhäuser zur Einhaltung dieser Sicherheitsstandards verpflichtet (§ 75c SGB V). Auch die Telematikinfrastruktur, das gesundheitssystemische Kommunikationsnetz und deren Anwendungen und mögliche datenanalytische Werkzeuge gelten als KRITIS-Elemente und sind u. a. IT-sicherheitstechnisch zertifizierungspflichtig (Kap. 4.1).

Die Kritikalitätsbewertungen von Daten, analytischen Verfahren und IT-Systemen sind vielschichtig und dynamisch. Unterschiedliche Faktoren, Risiken und mögliche Folgen werden berücksichtigt. Durch den technischen Fortschritt und durch gesellschaftliche Entwicklungen ergeben sich immer wieder neue Konstellationen, die situativ neue Lagebeurteilungen erfordern. Die Weiterentwicklung der Sicherheitsarchitektur vor allem kritischer Segmente ist ein kontinuierlicher Prozess. In Bereichen mit besonderer Kritikalität gewinnen normative Sicherheitsvorgaben sowie Zertifizierungs- oder Zulassungsverfahren kontinuierlich an Bedeutung.

Eigentumsrechte, Datenbesitz und Verfügungsgewalt

Eigentumsrechte sind die umfassendste alleinige Verfügungsgewalt, die natürlichen oder juristischen Personen über materielle oder immaterielle Dinge gewährt werden. Sie setzen die Existenz der jeweiligen Sache voraus und gehören in Deutschland zu den staatlich geschützten Grundrechten (Art. 14 GG). Eigentümer/innen können über ihre Sachen im Rahmen des geltenden Rechts weitgehend frei verfügen, Grenzen/Schranken sind gesetzlich zu definieren (z. B. die Beschränkung immaterieller Eigentumsrechte für Forschungsaktivitäten; Kap. 8.1). Für die Vergabe von Eigentumsrechten muss die Frage, wem die ausschließliche Verfügungsgewalt gewährt wird, im Rahmen des geltenden Rechts geklärt werden können. Sie ist insbesondere bei personenbezogenen Daten, an denen sowohl Personen als Datenobjekte als auch datenverarbeitende Stellen,

28 www.kritis.bund.de/SubSites/Kritis/DE/Einfuehrung/Sektoren/sektoren_node.html (10.11.2021)
29 www.dkgev.de/dkg/presse/details/bsi-gibt-gruenes-licht-fuer-dkg-sicherheitsstandard/ (10.11.2021)

die gewisse Leistungsschutz- und Geheimhaltungsrechte geltend machen können, offen. Dies ist ein Grund, warum ein Dateneigentum aus juristischer und ethischer Perspektive kritisch gesehen und derzeit mehrheitlich abgelehnt wird (DEK 2019, S. 18 ff.).

Durch die Verfügungsgewalt über die zur Datenerfassung und -speicherung erforderlichen Hard- und Softwarekomponenten sowie teilweise über definierte öffentliche Aufgaben oder über Einwilligungen kommen datenerhebende Stellen in den Besitz von (Roh-)Datenbeständen, müssen sie ggf. schützen und haben gewisse Nutzungsrechte. In gewerblichen Strukturen können diese Daten bisher weitgehend als Geschäftsgeheimnis[30] aufgefasst werden. Verpflichtungen zu deren Offenlegung müssen gesetzlich definiert werden. Private Unternehmen haben ein Grundrecht auf Berufsfreiheit und können im Rahmen des geltenden Rechts frei entscheiden, wie sie die Daten im Rahmen ihrer Nutzungsrechte u. a. für Data-Mining-Aktivitäten weiterverwenden, wie sie ihre Datenbestände anreichern und erweitern, welche Informationen sie ableiten, welche Algorithmen und analytischen Werkzeuge sie entwickeln und wie sie diese Ergebnisse verwerten. Sie können sie unternehmensintern einsetzen, Dritten Informationsdienstleistungen anbieten und/oder die analytischen Werkzeuge als Softwarebestandteile oder -produkte vermarkten. Im letzten Fall erreichen diese datenanalytischen Werkzeuge einen Produktstatus. Herstellende sind für deren Sicherheit und Leistung verantwortlich und haften ggf. bei Schäden (Kap. 3.4.2).

In öffentlichen Aufgabenbereichen regeln vielfältige Gesetze des öffentlichen Rechts einrichtungsspezifische Pflichten und Möglichkeiten zur Datenerhebung, -bereitstellung und -verwendung (z. B. Kap. 3.2 u. 5).

In der rechtlichen Auseinandersetzung wird teilweise dafür plädiert, statt über Dateneigentum zu diskutieren vielmehr die Verfügungsgewalt oder Hoheit über Daten und deren bessere Zugänglichkeit und Nutzung in den Blick zu nehmen (DEK 2019, S. 104; Hornung 2018, S. 17 ff.). Die derzeitigen datenbezogenen Rechtsstrukturen ermöglichen monopolartige Stellungen bezüglich der Haltung und Nutzung von Daten sowohl für Unternehmen als auch für öffentliche Einrichtungen. Diese Monopole verzerren oder verhindern Wettbewerbsstrukturen im digitalen Markt. Dazu wird derzeit diskutiert inwiefern u. a. große Plattformbetreiber ihre Datenbestände gegen Entgelt anderen datenverarbeitenden Stellen zugänglich machen und Algorithmen/Verfahren zur Datenerhebung und -verarbeitung offenlegen müssen. Zudem soll die Interoperabilität zwischen digitalen Diensten verbessert und neutrale Mittler zwischen datengebenden und analysierenden Personen etabliert werden. Die Datenmonopole öffentlicher

30 Ein Geschäftsgeheimnis ist jegliche Information unabhängig von der genauen Anordnung einzelner Bestandteile, die von wirtschaftlichem Wert und nicht allgemein bekannt ist (§ 2 Gesetz zum Schutz von Geschäftsgeheimnissen – GeschGehG)

Einrichtungen werden eher durch Transparenz- und Open-Data-Initiativen sowie die Errichtung spezifischer Datenrepositorien und -infrastrukturen überwunden (Kap. 3.2). Mehrere derzeit in Abstimmung befindliche europäische Verordnungen zielen darauf ab, Datenmonopole zu begrenzen, Wettbewerbsstrukturen digitaler Märkte zu stärken, Datenweiterverwendungsmöglichkeiten unter Achtung der Rechte Betroffener zu verbessern und Verbraucherrechte zu stärken (Kasten 3.2).

Kasten 3.2 Europäische Regulierungsinitiativen zur Verbesserung der Datennutzung

Das europäische *Gesetz über digitale Märkte* (EK 2020a) soll die Marktmacht sehr großer Plattformbetreiber begrenzen und die Wettbewerbsbedingungen für kleinere Unternehmen fairer gestalten (z. B. mittels Selbstbegünstigungsverboten, Datenzugangs- und Interoperabilitätsverpflichtungen) sowie die Wahlfreiheit für Endnutzer/innen stärken (sie sollen festlegen, welche Daten miteinander kombiniert werden dürfen, und nicht nur zustimmen). Die Verhandlungen zu diesem Gesetz wurden im März 2022 abgeschlossen. Es könnte Ende 2022 in Kraft treten.

Das europäische *Gesetz über digitale Dienste* (EK 2020b) definiert Verhaltensvorschriften für sich an Endverbraucher richtende Dienstleister, besonders strenge für sehr große Anbieter. Diese sollen u. a. Empfehlungen gebende oder Informationen vorsortierende Algorithmen in groben Zügen transparent machen und deren Risiken in Bezug auf Grundrechtsverletzungen von Meinungsfreiheit bis Diskriminierungsverbot (Kap. 3.3.1) jährlich bewerten. Sie sollen unter eine zentrale, bei der Europäischen Kommission angesiedelten Aufsicht gestellt werden. Die Verhandlungen zu diesem Gesetz wurden im April 2022 abgeschlossen. Nach einer Übergangsfrist könnte es 2024 in Kraft treten.

Mit dem *Daten-Governance-Gesetz* (EK 2020c) soll eine sichere, branchenübergreifende, europäische Dateninfrastruktur schaffen und die rechtssichere Datenweiterverwendung erleichtert werden. Ein Dateninnovationsrat soll Leitlinien und (Interoperabilitäts-)Standards entwickeln und die Datenportabilität voranbringen. Datentreuhänder, die große, teils auch geschützte Datenmengen verwalten und diese Dritten im Rahmen des geltenden Rechts teils gegen Entgelt bereitstellen, werden zur Neutralität verpflichtet. Datenaltruismus bzw. -spenden erhalten ein rechtliches Fundament. Das Gesetz könnte 2023 in Kraft treten.

Auch der 2022 vorgelegte Entwurf eines europäischen *Datengesetzes* (EK 2022) soll die bestehende Marktmacht sehr großer Plattform- bzw. Cloudanbieter begrenzen sowie den Zugang und die Nutzung der Daten für

kleinere Unternehmen verbessern. Unfaire, Machtasymmetrien begünstigende Vertragsklauseln sollen verboten, offene Standards und Schnittstellen perspektivisch verpflichtend vorgeschrieben werden. Cloudanbieter sollen einerseits verpflichtet werden, Daten vor dem Zugriff durch Drittländer zu schützen. Andererseits sollen staatliche Stellen innerhalb Europas in besonderen Situationen (z. B. pandemischen Lagen) bei besonderem Datenbedarf für öffentliche Zwecke vereinfachten Datenzugang erhalten können. Es zeichnet sich derzeit noch nicht ab, wann und in welcher Form dieser Gesetzesvorschlag in Kraft treten wird (Stand Mai 2022).

3.2 Umgang mit nichtpersonenbezogenen Daten: Beispiel Geodaten und nationale Geodateninfrastruktur

Vielfältige Sachverhalte u. a. zu Industrie-, Verwaltungs-, Gesellschafts- oder Umweltprozessen werden anhand von Datensätzen abgebildet, die keine direkten Bezüge zu einzelnen Personen haben. Der Zugang zu derartigen Daten ist zum einen von unterschiedlichen bereichsspezifischen gesetzlichen Vorgaben und zum anderen vom jeweiligen Geschäftsmodell der datenverarbeitenden Stelle und deren Auffassung in Bezug auf alleinige Datenverwendung oder Datenoffenlegung bestimmt. Bei Daten, die im Rahmen öffentlicher Aufgaben und unter Einsatz öffentlicher Mittel erhoben und gehalten werden, wird zunehmend Offenheit, Transparenz und Weiterverwendung gefordert. Durch die 2013 unter britischer Präsidentschaft von den G8-Ländern verabschiedete Open-Data-Charta bekennt sich auch Deutschland zur Datenoffenlegung (Kasten 3.3). Die zeitgleich verabschiedete Richtlinie 2013/37/EU[31] verpflichtet Deutschland in noch stärkerem Maße dazu.

Kasten 3.3 Open-Data-Konzepte

Open-Data-Konzepte sollen die Weiterverwendung von Datensätzen und -beständen erleichtern. Dazu wurden einerseits die für urheberrechtlich geschützte Werke entwickelten Creative-Commons-Lizenzen angepasst, mit denen Datensätze stufenweise offen zugänglich gemacht werden können: nur zur Ansicht; auch zur Weiterverwendung, nur nichtkommerziell oder auch zur kommerziellen Verwertung; mit oder ohne Quellenangaben. Andererseits hat die Organisation Open Knowledge International die in der freien

31 Richtlinie 2013/37/EU zur Änderung der Richtlinie 2003/98/EG über die Weiterverwendung von Informationen des öffentlichen Sektors

Softwareentwicklung existierenden Lizenzen zu Open-Database-Lizenzen (ODbL) weiterentwickelt, um Nutzenden unterschiedliche Freiheiten zu gewähren: zur Vervielfältigung; Weitergabe und Nutzung von Daten in der Ursprungsform; zur kreativen Verarbeitung, um Informationen abzuleiten oder neue Informationsdienste zu entwickeln (hier wäre Data-Mining zu verorten) sowie zur Transformation und Veränderung von Primärdaten, wodurch neue Datenbestände entstehen.

Open-Data-Konzepte sind weit mehr als die bloße Anpassung unterschiedlicher Lizenzmodelle. In der Realisierung verlangen sie in ihrer Maximalauslegung Primärdaten mit höchster Auflösung (nicht modifiziert), vollständig (alle Metadaten werden mitgeliefert), aktuell (zeitnah) und dauerhaft bereitzustellen, sie barrierefrei, d. h. maschinenlesbar und unter Verwendung offener Standards (keine spezifische Software erforderlich), zugänglich zu machen sowie die Weiterverwendung diskriminierungsfrei (für jede Person ohne Rechtfertigung), ohne Restriktionen (keine einschränkenden Nutzungsbedingungen), entgeltfrei mit Quellenangaben zu ermöglichen.

Open-Data-Konzepte werden vor allem bei Daten diskutiert, deren Erfassung und Speicherung einerseits im Rahmen öffentlicher Aufgaben oder mit öffentlichen Mitteln finanziert werden. Das sind vor allem Verwaltungsdaten (auch mit dem Begriff Open Government assoziiert). Andererseits wird der Begriff auch bei gemeinschaftlichen Datensammlungen angewendet (wie z. B. im Open-Street-Map-Projekt). Ausgenommen sind Daten, die rechtlichen Schutzmechanismen unterliegen. Das sind vor allem personenbezogene oder sicherheitsrelevante Daten sowie Daten, die Rechte Dritter (z. B. Geschäftsgeheimnisse, Urheberrechte) berühren (ausführlicher z. B. Bernsdorf et al. 2015, S. 132 f.; Dietrich 2011; Ksoll et al. 2017; Kuzev 2016). Open-Data-Konzepte erfordern erhebliche Anstrengungen zur Aufbereitung existierender Datenbestände sowie einen Kulturwandel in Bezug auf Informationsfreiheit und transparentem (Verwaltungs-)Handeln (Bundesregierung 2019d, S. 8).

Open-Data-Befürwortende betonen, dass es kosteneffizient sei, wenn vorhandene Daten umfangreich genutzt werden, und dass Mehrwert vor allem dann generiert werden könne, wenn möglichst viele Stellen vorhandene Daten vielfältig verwenden können (Grüebler 2014, S. 11; Ksoll et al. 2017). Diesen Mehrwert generiert in der Regel aber nicht die Stelle, die den zusätzlichen Aufwand für die Datenaufbereitung und -bereitstellung trägt, sondern Dritte, die diese Daten mit weiteren Daten verknüpfen, eigene Analysen durchführen und Informationsdienste entwickeln. Open-Data-Kritiker/innen weisen darauf hin, dass die Bereitstellung aktueller hochaufgelöster, qualitätsgesicherter Daten trotz Automatisierung oft mit erheblichem Aufwand verbunden sei, der bei Entgeltfreiheit nicht über klassische gewerbliche Geschäftsmodelle refinanziert werden könne. Dadurch würden wirtschaftliche

Anreize fehlen, quantitativ und qualitativ bessere Daten bereitzustellen. In Folge würden Wettbewerbsstrukturen verzerrt werden. Kritisiert wird auch, dass Datenfreigaben kaum kontrollierbar wären und Missbrauch begünstigen könnten (z. B. indem Personenbezüge hergestellt werden oder Daten manipuliert und dann fehlerhafte Informationen abgeleitet werden). Auch könne es Probleme bei Haftungsfragen geben (Kuzev 2016).

Ein beeindruckender Wandel von alleiniger Datennutzung zur Datenoffenlegung kann seit Jahren im Geodatenbereich beobachtet werden (z. B. Bernsdorf et al. 2015, S. 132; Gerlinger 2013, S. 12; TAB 2012, S. 74). Unterschiedliche Open-Data-Konzepte wurden in diesem Bereich entwickelt und realisiert. Erste Meilensteine aus den Vorreiterländern Großbritannien und USA sind (ausführlicher McKee 2012; Wangermann 2016):

- das 2004 in London initiierte Crowd-Sourcing-Projekt »OpenStreetMap«, bei dem viele Personen weltweit Geodaten erheben und zur freien Weiterverwendung bereitstellen, wobei auch die erforderliche Software zur Datenerhebung und -weiterverwendung offengelegt wird und sämtliche Aktivitäten von einer gemeinnützigen Stiftung orchestriert werden (ausführlicher z. B. Bernsdorf et al. 2015, S. 27 f.), sowie
- der von der US-amerikanischen Regierung 2008 beschlossene weltweit freie und kostenlose Zugang zu allen Erdfernerkundungsdaten des seit 1972 mit öffentlichen Mitteln finanzierten Satellitenprogramms »Landsat« (TAB 2012, S. 98 ff.), der ähnliche Entscheidungen u. a. in Europa beförderte (z. B. sind seit 2013 alle sicherheitsunkritischen Daten des europäischen Erdbeobachtungsprogramms »Copernicus« frei verfügbar).

Geodaten und nationales Geoinformationswesen

Geodaten sind alle Daten(sätze) mit direktem oder indirektem Bezug zu einem bestimmten Standort oder geografischen Gebiet (§ 3 GeoZG). Standorte werden über zwei- oder dreidimensionale Koordinatenangaben bestimmt (auch als primäre Metrik bezeichnet), geografische Gebiete über Kennziffern z. B. Postleitzahlen sowie Wahlbezirke eindeutig bezeichnet (sekundäre Metriken) (Bernsdorf et al. 2015, S. 40). Die Verknüpfung unterschiedlicher metrischer Ebenen erfolgt über spezifische Schlüsseltabellen, die den jeweiligen Kennziffern die Koordinaten der Gebietsgrenzen zuweisen und damit die Lage im Raum definieren. Maschinelle datenanalytische Konzepte einschließlich Data-Mining basieren auf zwei Datenteilbereichen (Bernsdorf et al. 2015, S. 40 ff.):

- *Geobasisdaten* beschreiben die Landschaft (Topografie) und die Liegenschaften (Grundstücke, Gebäude) zumindest einer Region. Sie sind die Basis für die Erstellung digitaler Landkarten bzw. Geländemodelle, die für

jegliche Verortung von Objekten und Geofachdaten, für raumbezogene Analysen (z. B. Hotspot-Analysen, Routenoptimierung) sowie zur Visualisierung von Analyseergebnissen erforderlich sind.

› *Geofachdaten* sind Datensätze aus jeglichen wirtschaftlichen, wissenschaftlichen oder gesellschaftlichen Bereichen, die einen eindeutigen Raumbezug haben, über den sie sowohl für raumbezogene Analysen nutzbar als auch auf digitalen Karten verortet werden können.

Die seit den 1990er Jahren kontinuierlich ausgebaute Normungsserie ISO 19100[32] ermöglicht eine hochgradig normierte Haltung jeglicher Geodaten durch standardisierte Daten- und Metadatenformate, Verknüpfungen unterschiedlicher Datensätze sowie die maschinelle Verarbeitung von Geodaten einschließlich raumbezogener Analysen. Die Anwendung dieser Normungsserie ist grundsätzlich freiwillig, sie kann jedoch gesetzlich vorgeschrieben werden z. B. zur Erfüllung öffentlicher Aufgaben. Die nachträgliche Normierung umfangreicher Bestandsdaten und deren detaillierte Beschreibung mit standardisierten Metadaten ist mit erheblichem Aufwand verbunden.

Die Begriffe Geodaten und Geoinformationen haben normativ eine erhebliche Schnittmenge: Geodaten repräsentieren Geoinformationen. Aufgrund der normierten Darstellung/Codierung sind Geodaten zwar maschinenlesbar, jedoch für den Menschen inhaltlich kaum unmittelbar erfass- und verarbeitbar. Die zur Datenverarbeitung und Informationserfassung notwendigen Hilfsmittel sind die bereits erwähnten *Geoinformationssysteme* (GIS) – spezielle Software, mit der Geodaten verarbeitet sowie Daten und Analyseergebnisse alphanumerisch oder grafisch dargestellt werden können (Bill/Fritsch 1991). Sowohl die Datenrepositorien als auch darstellende Landkarten gelten als Datenbanken, die vom urheberrechtlichen Leistungsschutz erfasst werden (Kap. 8.1).[33]

Geobasisdaten können vor Ort (in situ) direkt gemessen (Landesvermessung) oder anhand von Luft- oder Satellitenbildern ermittelt werden (Fotogrammetrie, Erdfernerkundung).[34] Die Erhebung von Geobasisdaten und deren dauerhafte Haltung ist in Deutschland eine öffentliche Aufgabe, die föderal aufgeteilt ist und wesentlich im Zuständigkeitsbereich der Länder liegt. Sie wird durch die jeweiligen Vermessungs- und Katastergesetze der Länder spezifiziert und durch das *amtliche Vermessungswesen* realisiert (Kummer et al. 2006). Zu diesem gehören vor allem die Vermessungs- und Katasterämter der Bundesländer sowie öffentlich bestellte, zertifizierte Vermessungsingenieur/innen (gilt als

32 Das europäische Komitee und das Deutsche Institut für Normung haben weite Teile der Serie als EN ISO 19100 bzw. als DIN EN ISO 19100 übernommen.

33 Landgericht München, Urteil vom 9.11.2005, Az. 21 O 7402/02 (Datenbankschutz für topografische Landkarten)

34 Die technischen Besonderheiten bei der Erhebung und bei der Verarbeitung von Satellitendaten und deren Analysepotenziale vor allem im Kontext der Entwicklungszusammenarbeit werden TAB-Arbeitsbericht Nr. 154 ausführlich thematisiert (TAB 2012).

freier Beruf zur Erbringung von Dienstleistungen höherer Art).[35] Letztere vermessen räumlicher Gegebenheiten vor Ort und zunehmend auch aus der Luft im Auftrag der jeweiligen Ämter und übermitteln ihre erhobenen Daten an diese. Vermessungs- und Katasterämter führen Geodaten in unterschiedlichen Formaten (Dokumente, Tabellen, Luftbilder) zusammen und pflegen die amtlichen Geobasisdatenbestände in spezifischen Repositorien, auf die unterschiedliche Geoinformationssysteme aufbauen (Kasten 3.4). Zum Vermessungswesen gehören auch einige Ämter auf Bundesebene, darunter das Bundesamt für Kartographie und Geodäsie. Es pflegt die amtlichen nationalen Referenzdatenbestände sowie die darauf aufbauenden digitalen Landkarten und Raummodelle.[36] Vor allem die Landeseinrichtungen unterliegen traditionell erheblichen Selbstfinanzierungserfordernissen und arbeiten bisher weitgehend gebührenfinanziert.

Kasten 3.4 Amtliche Geoinformationssysteme ATKIS und ALKIS

Das amtliche topographisch-kartographische Informationssystem (ATKIS) enthält primäre topografische Geobasisdaten Deutschlands sowie zentrale Schlüsseltabellen, die geografische Objekte und Flächen anhand ihrer Lagepunkte (Koordinaten der Flächenränder) definieren, diese eindeutig bezeichnen und signieren (labeln) und damit u. a. Siedlungsflächen, Verkehrsnetze, Vegetation, Gewässer, Relief und Verwaltungsgebiete ausweisen. Diese Schlüsseltabellen ermöglichen die Verknüpfung mit diversen Datenbanken, die weitere Daten zu den jeweiligen Flächen oder Objekten halten. Diese Geobasisdaten und Schlüsseltabellen bilden das *zentrale amtliche Referenzmodell* Deutschlands – quasi einen digitalen Zwilling des realen Geländes. Das Geoinformationssystem ATKIS ermöglicht die kontinuierliche Aktualisierung der zentralen amtlichen Datenbasis, die Darstellung der Daten anhand digitaler Karten und deren Bereitstellung für weitere Anwendungen mittels spezifischer Dienste.

Das amtliche Liegenschaftskatasterinformationssystem (ALKIS) basiert auf einer digitalen Liegenschaftskarte, deren zentrale Schlüsseltabelle alle Liegenschaften/Flurstücke als geografische Gebiete definieren und anhand der Flurstückcodierung eindeutig bezeichnen. Diese Codierung bildet die Schnittstelle u. a. zum automatisierten Liegenschaftsbuch – der Datenbank, die alle amtlichen Eintragungen laut Grundbuchordnung (GBO) für jedes

35 Freie Berufe haben berufsspezifischen Standesregeln, die einerseits das Berufsverständnis fundieren sollen und andererseits bei Nichteinhaltung berufsrechtliche Konsequenzen haben können. In den Standesregeln von Vermessungsingenieur/innen wird u. a. die unabhängige und sorgfältige Berufsausübung im Bewusstsein der Bedeutung der Daten betont und die Schweigepflicht auch berufsrechtlich verankert (www.bdvi.de/application/files/1915/6093/1809/standesregeln_bdvi_06062009.pdf; 10.11.2021).

36 Rechtsgrundlage ist das Gesetz über die geodätischen Referenzsysteme, -netze und geotopographischen Referenzdaten des Bundes (Bundesgeoreferenzdatengesetz – BGeoRG).

Flurstück enthält (u. a. Eigentumseintrag, Nutzungsart, Grundschulden). Grundbucheintragungen bilden das juristische Fundament des Eigentums an Grund und Boden. Einzelne Liegenschaftsdatensätze haben aufgrund der Eigentumseinträge einen Personenbezug.

Die Art und Weise der Datenbereitstellung werden durch Geodatenzugangsgesetze auf Bundes- und Landesebenen definiert. Diese verpflichten die jeweiligen Ämter Geo(basis)daten in ihrem Verantwortungsbereich ohne besondere Kritikalität für die Weiterverwendung öffentlich zugänglich zu machen. Die Ämter gelten als Dateneigner, die datenverarbeitenden Stellen mittels Lizenzen definierte Nutzungsrechte gewähren (teilweise gegen Gebühr). Letztere müssen bei der Datenverwendung Quellen erkennbar vermerken. Damit werden die Strukturen des Urheberrechts (Leistungsschutz) auf den öffentlichen Geodatenbereich übertragen. Die technische und rechtliche Realisierung des Datenzugangs erfolgt zunehmend über die *Geodateninfrastruktur* (GDI).

Für die Georeferenzierung und Haltung von Fachdaten anderer öffentlicher Aufgabenbereiche sind vielfältige Fachressorteinrichtungen auf Landes- und Bundesebene zuständig. Dadurch wird die Georeferenzierung von Fachdaten zu einer Querschnittsaufgabe der öffentlichen Verwaltung. Unter den Fachressorts nimmt das Umweltressort eine besondere Stellung ein. Zum einen haben die meisten Umweltdaten per se Raumbezüge. Zum zweiten ist es der Bereich, dessen Ressorteinrichtungen seit vielen Jahren durch Umweltinformationsgesetze auf Bundes- und Landesebene zu einer verständlichen, aktuellen, weitgehend entgeltfreien Information der Öffentlichkeit gesetzlich verpflichtet sind und die in Folge seit Jahren Umweltinformationssysteme ausbauen – fachspezifische GIS, bei denen grafische Darstellungen im Rahmen von Berichts- und Abfragesystemen eine wichtige Rolle spielen und die dafür auf die nationalen Referenzdatenmodelle zugreifen. Alle öffentlichen Facheinrichtungen, die Geobasis- oder Geofachdaten verarbeiten, bilden gemeinsam das *Geoinformationswesen*.

Mit der Entwicklung der Fernerkundungstechnologie sowie der Ortungsfähigkeit mobiler Geräte und mit der zunehmenden Verfügbarkeit von Geodaten können auch Unternehmen vielfältige Geodaten erheben oder von anderen beziehen, um digitale Karten und Raummodelle zu erstellen, Fachdaten zu verorten, raumbezogene Data-Mining-Projekte durchzuführen und/oder geodatenbasierte Anwendungen/Informationsdienste zu entwickeln. Derartige Unternehmen werden auch als *Geoinformationswirtschaft* oder GeoIT-Branche bezeichnet (ausführlicher u. a. Bernsdorf et al. 2015, S. 22 ff.).

Personensensibilität und Sicherheitskritikalität von Geodaten

Aufgrund der zunehmenden technischen Möglichkeiten können raumbezogene Daten kontinuierlich detailgenauer erfasst und über unterschiedliche Bezüge

vielfältig verknüpft und erweitert werden. Dadurch können Geodaten sowohl unmittelbar als auch durch die Verknüpfung mit anderen Daten und durch Analyse u. a. besondere Liegenschaften und dortige Vorgänge offenbaren, Personen identifizieren oder orten[37] und damit Grundrechte von Personen (Privatheit, Geschäftsgeheimnisse, Eigentum) verletzen oder bezüglich der inneren und äußeren Sicherheit bzw. der Gefahrenabwehr als kritisch bewertet werden. Die Erhebung und/oder Weitergabe kritischer Geodaten soll auf unterschiedliche Weise beschränkt und kontrolliert werden:

Öffentliche Einrichtungen des Geoinformationswesens dürfen im Rahmen ihrer gesetzlich definierten Aufgaben auch personensensible oder sicherheitskritische Geodaten erheben und halten. Der Zugang für Dritte über die Geodateninfrastruktur wird jedoch beschränkt, wenn (§ 12 GeoZG)

- schutzwürdige Interessen von Personen beeinträchtigt werden (Recht auf Privatheit bzw. Betriebs-/Geschäftsgeheimnisse, Rechte am geistigen Eigentum), es sei denn, betroffene Personen haben zugestimmt;
- durch den Datenzugang laufende Ermittlungen oder Gerichtsverfahren oder
- bedeutsame Schutzgüter der öffentlichen Sicherheit, die Verteidigung oder internationale Beziehungen gefährdet werden.

Da Unternehmen und Organisationen des privaten Rechts jedoch zunehmend selbst hochaufgelöste Geodaten erheben und gewerblich verwerten können, sind sie auf einen Zugang zu amtlichen Referenzdaten immer weniger angewiesen. Mit diversen ortungsfähigen Geräten können sie bodennah räumliche Gegebenheiten abbilden und vermessen oder Aktivitätsmuster von gerätetragenden Personen georeferenzieren und aufzeichnen. Die Grenzen der Zulässigkeit bodennaher Datenerfassungen wurden bisher vor allem in Bezug auf die Achtung der Grundrechte von Personen durch Rechtsprechung konkretisiert und beziehen sich besonders auf georeferenzierte Abbildungen.[38] Potenziell kritische Geodaten können auch aus der Luft mittels Drohnen, Flugzeugen oder Satelliten und deren spezifischen Sensor- oder Kamerasystemen erhoben werden. Um den

37 Eine frühere ausführliche Thematisierung gesellschaftlicher Herausforderungen im Kontext von Ortungsdaten bietet z. B. Hilty et al. (2012).

38 Beispielhaft seien gerichtliche Klärungen im Kontext von Google Street View genannt, die bei georeferenzierten Bildern immer wieder die Grenzen zwischen Urhebenden und deren Berufs- und Panoramafreiheiten (Recht von Positionen des öffentlichen Raumes aus Liegenschaften zu fotografieren) und allgemeinen Persönlichkeitsrechten und Geheimhaltung definieren. So dürfen in Deutschland mit Kameras Gebäude und private Grundstücke nur aus einer Höhe von maximal 2,9 m aufgenommen werden (Grenzwerte variieren z. T. zwischen einzelnen Ländern, in der Schweiz liegt er z. B. bei 2 m). Auf Abbildungen müssen Personen und Autokennzeichen unkenntlich gemacht werden, bei Fassadenaufnahmen dürfen Gebäudeeigentümer/innen ebenfalls die Anonymisierung durch Vergröberung bzw. Verpixelung verlangen (wobei bei Gebäuden mit mehreren Eigentumsparteien und Mietverhältnissen komplexere Rechtssituationen entstehen können) (ausführlicher z. B. Ernst 2010).

Einsatz hochwertiger Fluggeräte und die Erhebung potenziell kritischer Geodaten zu kontrollieren, ist zum einen eine Betriebserlaubnis erforderlich (für Drohnen ab 5 kg, darunter reicht ein Kenntnisnachweis [Abschnitt 5a LuftVO[39]). Zum anderen gibt es Überflug- damit auch Datenerhebungsverbote mit Drohnen über privaten Grundstücken und Gebäuden (mit Erlaubnisvorbehalt durch Betroffene)[40] sowie in und über sensiblen Situationen und Bereichen (u. a. über Menschenansammlungen sowie Einsatzorten von Sicherheits- und Rettungskräften, über kritischen Infrastrukturanlagen sowie öffentlichen Einrichtungen und Industrieanlagen). Auch der Betrieb hochwertiger Erdfernerkundungssysteme (Satelliten und Bodenstationen) sowie die Erhebung und Weitergabe hochaufgelöster aktueller Satellitendaten sind nach dem Satellitendatensicherheitsgesetz (SatDSiG)[41] erlaubnispflichtig. Satellitenbetreibende und -datenanbietende Stellen müssen die Kritikalität entsprechender Datenanfragen in einem zweistufigen Verfahren prüfen. Zuerst prüfen sie selbst den Informationsgehalt der gewünschten Daten und die Zuverlässigkeit der beantragenden Stelle, dann holen sie eine behördliche Bestätigung ein (§§ 17 ff. SatDSiG).

Auch wenn es für die Erfassung und Bereitstellung von Geodaten einen Rechtsrahmen gibt, sind sowohl die Bewertung der Sicherheitskritikalität oder Personensensitivität von Geodaten als auch die entsprechende Regulierung der Vorgehensweisen kontinuierliche Herausforderungen, denn dafür sollten nicht nur die Auflösung und die Aktualität der jeweiligen Datenerfassungen berücksichtigt werden, sondern auch die für eine datenverarbeitende Stelle verfügbaren weiteren Datenbestände und die Leistungsfähigkeit der datenanalytischen Verfahren, die datenverwendende Stellen einsetzen können.

Einen spezifischen Rechtrahmen für die Analyse von Geodaten einschließlich Data-Mining gibt es bisher nicht. Laut Bernsdorf et al. (2015, S. 213) zeigen Diskussionen unter Geodienstleistern aus Forschung und Wirtschaft sowie entsprechender Verbandsaktivitäten, dass es bezüglich der Bewertung der Kritikalität von Geodaten und deren Verarbeitung sowohl unterschiedliche Positionen als auch Unsicherheiten bezüglich der Grenzen der zulässigen Datenverarbeitung gibt. Diejenigen, die Geodaten zur Entwicklung neuer Informationsdienste nutzen wollen, bewerten die Kritikalität von (Geo-)Daten tendenziell eher schwächer und fordern eher deren Zugänglichkeit. Diejenigen, die Personen mit ihren Grundrechten oder diverse sicherheitsrelevante Dinge schützen wollen

39 Luftverkehrs-Ordnung (LuftVO) (BGBl. I S. 1894)

40 Die Abbildung von Grundstücken aus der Luft gilt als Eingriff in das allgemeine Persönlichkeitsrecht, wenn private Lebensbereiche gezeigt werden, die von öffentlichen Plätzen nicht einsehbar sind, auch wenn keine Personen abgebildet sind (BVerfG, Beschluss vom 2.5.2006, 1 BvR 507/01)

41 Gesetz zum Schutz vor Gefährdung der Sicherheit der Bundesrepublik Deutschland durch das Verbreiten von hochwertigen Erdfernerkundungsdaten (Satellitendatensicherheitsgesetz – SatDSiG)

(von kritischen Infrastrukturen bis zu militärischen Objekten), fordern eher einen restriktiveren Umgang mit Geodaten, der jedoch ins Leere läuft, je mehr hochauflösende Geodaten aus unterschiedlichen Quellen weltweit verfügbar werden (ausführlicher z. B. Bernsdorf et al. 2015, S. 180 ff.).

Auf- und Ausbau der Geodateninfrastruktur Deutschland

Die Geodateninfrastruktur Deutschland (GDI-DE) soll Geodaten aus unterschiedlichen Quellen für die Weiterverwendung rechtssicher zugänglich machen (Bernsdorf et al. 2015, S. 18 ff.). Sie hat folgende technische Komponenten:

- dezentrale Geodatenrepositorien in der Verantwortung datenbereitstellender Einrichtungen, die unter Anwendung von Geodatennormen und interoperablen Standards Geodatensätze halten und mittels standardisierter Metadatensätze katalogisieren, sodass Datensätze ähnlich wie Texte über Bibliothekskataloge gesucht und bereitgestellt werden können;
- ein Netzwerk, das die Repositorien der sich beteiligenden Einrichtungen verknüpft sowie
- Zugangsportale zu den Repositorien und deren Katalogen mit spezifischen Diensten
  - zur Datensuche (greifen auf die Metadatenfiles der Kataloge zu),
  - zur Visualisierung (stellen Geodaten anhand von Karten grafisch dar),
  - zur Transformation (u. a. Umrechnungen, Maßstabsveränderungen),
  - zum Datendownload sowie
  - zur digitalen Abwicklung des damit verbundenen Geschäftsverkehrs (Nutzerregistrierung, Lizenzerteilung, Bezahlung).

Die USA förderten bereits in den 1990er Jahren den Aufbau von Geodateninfrastrukturen, um den Zugang zu Geodaten, die mit öffentlichen Mitteln erhoben wurden, zu vereinfachen und sie leichter weiterverwenden zu können. Dieses Engagement wurde flankiert von den Aktivitäten zur Kommerzialisierung vor allem des Satellitendatenbereichs, die den Betrieb von Satelliten und die Erhebung von Geodaten auch für Unternehmen ermöglichte und die das staatliche Monopol der Geodatenerfassung beendete (ausführlicher TAB 2012, S. 96 ff.).

Auch in Deutschland gab es bereits vor der Jahrtausendwende erste Schritte, um amtliche Geodaten zumindest für öffentliche Aufgaben besser nutzen zu können: Das Bundesamt für Kartographie und Geodäsie (BKG) begann mit dem Aufbau eines eigenen Geodatenzentrums, in dem ausgewählte Geobasisdaten der Länder auf Bundesebene zusammengeführt wurden, um sie anderen Fachressorteinrichtungen der Bundesebene für deren jeweilige Aufgaben bereitzustellen. Die Bundeseinrichtungen vernetzten sich im interministeriellen Ausschuss für Geoinformationswesen (IMAGI), um ihren Geodatenbedarf und ihr

Vorgehen abzustimmen. Diesen Aktivitäten und dem Geodatenbedarf auf Bundesebene standen vielfältige datenerhebende öffentliche Einrichtungen als Datenhalter mit Selbstfinanzierungserfordernissen auf kommunaler und Landesebene gegenüber.

Ein Wandel im Verwaltungsdenken bezüglich des Zugangs und der Weiterverwendung amtlicher Geo(basis)daten setzte in Deutschland nach der Jahrtausendwende ein. Befördert wurde er durch die entstehenden internationalen Wettbewerbsstrukturen der Geodatenanbieter, eine sich aus wissenschaftlichen Einrichtungen herauslösende, eigenständig etablierende GeoIT-Branche, immer umfangreichere informationstechnische Möglichkeiten und euphorische Marktprognosen bezüglich der Entwicklung neuer geodatenbasierter Services und damit möglicher Mehrwertgenerierung. Laut Bernsdorf et al. (2015, S. 25 ff.) könne man im Geodatenbereich seitdem beobachten, wie sich informationstechnische und normative Entwicklungsschritte sowie öffentliches und privatwirtschaftliches Engagement für einen strukturierten Zugang zu Geodatenbeständen in öffentlicher Hand wechselseitig verstärken, auch wenn es im Detail unterschiedliche Positionen, Rahmenbedingungen und Logiken bei beteiligten Akteuren gibt. Eine Art Goldgräberstimmung sei entstanden – auch bei der Schaffung gesetzlicher Rahmenbedingungen für eine breitere Nutzung von Geodaten der öffentlichen Hand. Einerseits wollte man den Markt nicht nur ausländischen Anbietern überlassen und suchte Möglichkeiten, amtliche Geodatenbestände ebenfalls wirtschaftlich nutzbar zu machen. Andererseits sollten dem prognostizierten prosperierenden Markt auch gewisse Regeln gegeben werden. Aufgrund verteilter Verantwortlichkeiten sind dafür vielfältige Abstimmungen und gesetzliche Aktivitäten auf europäischer, Bundes- und Landesebene erforderlich (Kasten 3.5).

Kasten 3.5 Rechtsgrundlagen der Geodateninfrastruktur

Die Richtlinie 2007/2/EG[42] zielt darauf ab, EU-weit harmonisierte, nationale Geodateninfrastrukturen aufzubauen, um diese insbesondere für eine gemeinschaftliche Umweltpolitik nutzen zu können. Sie verpflichtet die Mitgliedstaaten Geobasisdaten und definierte Geofachdaten mit Umweltbezug zunehmend standardisiert bereitzustellen. Geodatenzugangsgesetze und deren Rechtsverordnungen auf Bundes- und Landesebene setzen diese Richtlinie national um und definieren für die jeweiligen datenhaltenden Ämter

- die Gültigkeit des Urheberrechtsgesetzes (Ämter gelten teilweise als Dateneigner, dürfen Nutzungsrechte mittels Lizenzen vergeben und Gebühren erheben);

42 Richtlinie 2007/2/EG zur Schaffung einer Geodateninfrastruktur in der Europäischen Gemeinschaft (INSPIRE)

- die Bereitstellung standardisierter Geodaten und Metadatenfiles (Verwendung Normungsserie EN ISO 19100, Maschinenlesbarkeit, Dauerhaftigkeit, Aktualität) sowie definierter Dienste zur Datenverarbeitung und zum Geschäftsverkehr;
- die Weiterverwendung, wobei die GeoZG der meisten Bundesländer zwischen öffentlichen Aufgaben, anderer nichtkommerzieller und kommerzieller Nutzung differenzieren und letztere begrenzen;
- Nutzungsbeschränkungen, um Grundrechte von Personen oder bedeutende Schutzgüter der öffentlichen Sicherheit (kritische Infrastrukturen), Verteidigung sowie internationale Beziehungen nicht zu gefährden sowie
- Haftungsbeschränkungen und -ausschlüsse (keine Sachschäden, bei Fahrlässigkeit).

Das Bundesgeoreferenzdatengesetz (BGeoRG) definiert die Aufgaben des Bundesamtes für Kartographie und Geodäsie bezüglich der Bereitstellung neutraler Geobasisdaten und nationaler Referenzsysteme, digitaler Karten und Raummodelle. Zugang ohne Gegenleistung erhalten Bundeseinrichtungen zur Wahrnehmung öffentlicher Aufgaben.

Nach Diskussionen und Beschlussfassungen u. a. auch im Deutschen Bundestag begann 2003 das gemeinschaftliche Vorgehen von Bund, Ländern und Kommunen zum Aufbau der Geodateninfrastruktur Deutschland (Bundesregierung 2003; SPD/BÜNDNIS 90/DIE GRÜNEN 2001). Seitdem fordert das Parlament in jedem 3. Jahr der Legislaturperiode einen Fortschrittsbericht zur Entwicklung des Geoinformationswesens von der Regierung ein. Diese Berichte dokumentieren Herausforderungen beim Auf- und Ausbau der GDI-DE, erreichte Meilensteine sowie definierte neue Etappen und Ziele (Bundesregierung 2003, 2005, 2008, 2012, 2017 u. 2021b). Beim Bundesministerium des Innern (BMI) wurde eine Lenkungsgruppe eingerichtet, die alle GDI-DE-Aktivitäten koordiniert und für die Berichterstattung verantwortlich ist. In der Anfangsphase stand das technische Datenmanagement im Fokus. Parallel dazu mussten vielfältige rechtliche Fragen geklärt werden.

Die Harmonisierung und Vereinfachung diverser kommunaler-, landes-, bundes- und aufgabenspezifischer Zugriffsregeln, Nutzungsbedingungen und Kostenmodelle war und ist ein Entwicklungsprozess mit vielen Etappen (Bernsdorf et al. 2015, S. 170 ff.). 2005 verständigten sich die Bundes- und Landeseinrichtungen mit den Entgeltrichtlinien auf drei deutschlandweit gültige Nutzungskategorien (öffentliche Aufgaben, Forschung und gewerbliche Nutzung), hielten jedoch an ihren Verwertungsrechten fest, was mit komplizierten Lizenzverfahren einherging. Innerhalb der Bundesverwaltung verständigte man sich auf eine entgeltfreie Nutzung von Geodaten und -diensten für gesetzlich definierte Aufgaben, sofern keine wirtschaftliche Weiterverwendung erfolgte. Im

nächsten Schritt wurden Lizenzierungsverfahren digitalisiert und Abrechnungsprozesse verwaltungsübergreifend vereinheitlicht. 2008 ging man zu bundeseinheitlichen Gebührenkatalogen über und einigte sich darauf, bei der Gebührenbemessung nur die Kosten für die Bereitstellung, nicht aber die für die Datenerhebung zu berücksichtigen.

Seit 2013 stellen Bundesbehörden ihre selbst erhobenen Geodaten, deren Metadaten sowie ihre Dienste zur Suche, Darstellung oder zum Download für jegliche Weiterverwendung kostenlos bereit (Quellenangaben werden gefordert, auf Nutzungslizenzen hingegen verzichtet). Jedoch werden vor allem hochaufgelöste Geodaten nicht auf Bundesebene, sondern von Landes- oder kommunalen Einrichtungen erhoben. Dieser bisher weitreichendsten Open-Data-Initiative des Bundes schlossen sich 6 Bundesländer an.[43] In anderen Bundesländern werden nur Such- und Darstellungsdienste kostenlos bereitgestellt, beim eigentlichen Datendownload wird an Nutzungslizenzen und Gebührenerhebung festgehalten. Dadurch sehen sich lizenznehmende Stellen trotz einiger Vereinfachungen nach wie vor mit komplexen Strukturen konfrontiert, denn die Länder können diverse Positionen der Lizenzvereinbarungen nach wie vor individuell gestalten. Laut Bernsdorf et al. (2015, S. 199 f.) ist diese Situation in Kombination mit jährlichen Gebührenanpassungen für lizenznehmende Stellen nach wie vor kompliziert und bietet kaum längerfristige Planungssicherheit.

Auch die Kritikalität insbesondere von hochaufgelösten Geodaten wurde thematisiert, um Grenzen für die Datenzugänglichkeit über die GDI festzulegen (u. a. Karg 2008). Bisher werden über die GDI nur unkritische Geodaten bereitgestellt. Eine Instanz, die ggf. die Zuverlässigkeit datenverarbeitender Stellen oder einzelner Analyseanträge prüft und ggf. hochaufgelöste Geodaten situativ für Einzelanalysen freigibt, ist bisher nicht in die GDI eingebunden.

Der GDI-Aufbau wurde nach 10 Jahren evaluiert (Bundesregierung 2012a). Nicht alle ursprünglich gesetzten Zeitziele wurden erreicht, etliche Erfolge blieben hinter Notwendigkeiten zurück. Als Gründe wurden die Vielzahl beteiligter Akteure, die begrenzte Ressourcenausstattung und eine geringe Aufgabenpriorisierung genannt. Mit Blick nach vorn wurde eine Bedarfs- und Nutzungsprüfung bereitgestellter Daten und ein Vergleich mit anderen Datenanbietern empfohlen, um öffentliche Angebote auf ihre Relevanz zu prüfen. Eine Nationale Geoinformationsstrategie (NGIS) solle sich perspektivisch stärker darauf konzentrieren, welche Aufgaben zur staatlichen Grundversorgung gehören und welche von privatwirtschaftlichen Akteuren übernommen werden können. Diese Strategie wurde 2015 verabschiedet und beschreibt Aufgabenschwerpunkte und Ziele für die nächsten 10 Jahren (AG NGIS 2015). Die Schaffung und zuverlässige Bereitstellung einer nationalen Geodatenbasis (einschließlich Metadatenbasis) wurde als staatliche Aufgabe mit geteilter Zuständigkeit von

43 Berlin, Brandenburg, Hamburg, Nordrhein-Westphalen, Sachsen, Thüringen

Bund, Ländern und Kommunen bestätigt. Die Entwicklungsziele sind auf die Intensivierung der Nutzung ausgerichtet, explizit auf die

- Mehrfachnutzung der nationalen Geodaten;
- Senkung der nach wie vor bestehenden nutzungsrechtlichen Hemmnisse (u. a. sollen Lizenzen nutzerfreundlicher werden, Kostenstrukturen vereinheitlicht sowie Haftungsfragen in Bezug auf die Richtigkeit bereitgestellter Daten und die rechtskonforme Verarbeitung unter Einhaltung von Datenschutz und Datensicherheit das Geben und Nehmen klarer regeln) sowie
- Förderung von Innovationen, wobei Staat, Wissenschaft und Wirtschaft gleichermaßen verantwortlich seien, neue Angebote anzustoßen, zu erproben und die Implementierung zu unterstützen (u. a. sollen geeignete Plattformen zur Kommunikation und Koordination gemeinsamer Projekte gezielt ausgebaut werden).

Die stärkere Ausrichtung auf die Datennutzung schlägt eine stärkere Brücke zu Data-Mining-Aktivitäten. Auch wenn die aufwendigen Abstimmungsprozesse und die nach wie vor bestehenden Nutzungshemmnisse im Geodatenbereich mitunter kritisiert werden, gehört Deutschland bezüglich des GDI-Aufbaus und der Implementierung der Richtlinie 2007/2/EG im europäischen Vergleich zu den Vorreiterländern (Cetl et al. 2017). Von den Erfahrungen beim GDI-Aufbau könnten ggf. auch andere öffentliche Bereiche profitieren, die ebenfalls beginnen, Dateninfrastrukturen auf- und auszubauen.

Bereitstellung anderer Verwaltungsdaten

Nur wenig zeitversetzt zum Geodatenbereich starteten auch für andere öffentliche Bereiche Initiativen für mehr Offenheit bei den im Rahmen öffentlicher Aufgaben generierten Daten und Informationen. Gesetze auf Bundes- und Landesebene wurden erlassen, es gab jedoch keinen mit dem Geodatenbereich vergleichbaren Enthusiasmus, Verwaltungsdaten allgemein öffentlich zugänglich zu machen.

Die ab 2006 in Kraft getretenen *Informationsfreiheitsgesetze* auf Bundes- und Landesebenen attestieren zwar einen voraussetzungslosen Anspruch für jedermann auf Zugang zu Informationen, die öffentliche Einrichtungen im Rahmen ihrer Aufgaben gewinnen. Darunter fallen nicht nur Daten, Tabellen, Karten und Bilder, sondern auch Analysen, Berichte, Schriften und sonstige Aufzeichnungen (keine Notizen, Entwürfe, Stichpunkte). Ausgenommen sind jedoch jegliche Informationen, die aus diversen Gründen geschützt sind (u. a. zum Schutz vielfältiger öffentlicher Belange und behördlicher Entscheidungsprozesse, informationeller Selbstbestimmungsrechte, von Geschäftsgeheimnissen oder geistigem Eigentum). Die jeweiligen öffentlichen Einrichtungen sind ver-

pflichtet, Daten- und Informationsbestände zu verzeichnen und diese Verzeichnisse elektronisch zugänglich und entgeltfrei nutzbar zu machen. Sie müssen die Richtigkeit der Daten und Informationen nicht immer prüfen. Laut Informationsfreiheitsgesetzen kann die Daten- und Informationsbereitstellung mündlich, schriftlich oder elektronisch erfolgen, teilweise dürfen Aufwandsgebühren in Rechnung gestellt werden.

Nach der Unterzeichnung der G8-Open-Data-Charta und infolge der Richtlinie 2013/37/EU wurden auf Bundes- und Landesebenen *E-Government-Gesetze* verabschiedet, die wesentlich auf elektronische Verwaltungsabläufe abzielen, aber auch die Bereitstellung maschinenlesbarer Verwaltungsdaten verbessern sollen. Datenbereitstellende Behörden sind bisher zu keinerlei Datenprüfung verpflichtet und übernehmen keine Haftung in Bezug auf Richtigkeit, Qualität, Aktualität und dauerhafte Bereitstellung der Daten. Der Aufbau von Dateninfrastrukturen wird bisher nicht explizit definiert. Dem Bundestag ist alle zwei Jahre über den Fortschritt zu berichten.

Diverse Evaluationen weisen seit Jahren darauf hin, dass die Bereitstellung und Weiterverwendung von Verwaltungsdaten nur schleppend vorankommen. Öffentliche Einrichtungen stellen kaum Ressourcen für die Offenlegung ihrer Verwaltungsdaten bereit (Bundesregierung 2016a, S. 7 ff. u. 2019a; WD 2019). Teilweise wird ihnen ein noch nicht ganz vollzogener Bewusstseinswandel attestiert (DEK 2019, S. 156), in Einzelfällen sogar eine Weigerung festgestellt, Informationspflichten zu erfüllen (BfDI 2020, S. 9). Im internationalen Vergleich gehört Deutschland bisher keinesfalls zu den Vorreitern offener (Verwaltungs-)Daten (Wangermann 2016).[44]

Die im Anschluss verabschiedeten *Informationsweiterverwendungsgesetze* zielen zwar auf die Intensivierung der maschinellen Weiterverwendung von Daten und Informationen – und sollten damit auch Data-Mining-Aktivitäten ermöglichen –, formulieren dann jedoch weitgehend unverbindliche Allgemeinaussagen: Öffentliche Einrichtungen sollen ihre ohnehin offenzulegenden Daten/Informationen und deren Metadaten/Beschreibungen möglichst elektronisch, in offenen und maschinenlesbaren Formaten bereitstellen, sofern das mit verhältnismäßigem Aufwand möglich ist. Sie können für die bereitzustellenden Daten/Informationen Nutzungslizenzen vergeben. Letztere sollen allgemein zugänglich sein, keine Ausschließlichkeitsvereinbarungen enthalten (Ausnahmen sind möglich) und jegliche Weiterverwendung gleichbehandeln (keine Differenzierung zwischen gewerblich und nichtgewerblich). Nutzungsentgelte dürfen in den meisten Bundesländern erhoben werden. Sofern Informationen be-

44 Die Chancen der Öffnung und Bereitstellung von Verwaltungsdaten werden in einem separaten Projekt des TAB thematisiert (Chancen der digitalen Verwaltung, www.tab-beim-bundestag.de/de/untersuchungen/u40200.html; 10.11.2021).

reits mit Metadaten beschrieben und in maschinenlesbaren Formaten über öffentliche Netze zugänglich sind, sollen diese Metadaten auch einem zentralen Webportal zur Verfügung gestellt werden (GovData.de).

Ende 2019 startete die Bundesregierung eine neue Initiative, um die verantwortungsvolle Datenbereitstellung und -nutzung in Deutschland signifikant zu steigern (Bundesregierung 2019c). 2021 verabschiedete sie zum einen die nationale Datenstrategie, mit der der Bund diesbezüglich zum Vorreiter werden will (Bundesregierung 2021a). Zum anderen wurde auf Bundesebene das E-Government-Gesetz[45] geändert und das Informationsweiterverwendungsgesetz durch das Gesetz zur Nutzung von Daten des öffentlichen Sektors ersetzt.[46] Dadurch werden öffentliche Stellen des Bundes als Datenbereitsteller umfangreicher benannt (auch Unternehmen mit öffentlich-rechtlichen Aufgaben und Forschungseinrichtungen), jedoch gibt es nach wie vor zahlreiche Ausnahmen, u. a. wird die Datenbereitstellung und -nutzung der sozialgesetzlich definierten Selbstverwaltungskörperschaften (darunter Einrichtungen der gesetzliche Krankenversicherung; Kap. 5) eigenständig reguliert. Die Datenbereitstellung erfolgt weitgehend im etablierten Rahmen (Metadaten sollen über das GovData-Portal bereitgestellt werden, die Datensätze über öffentlich zugängliche Netze, aller zwei Jahre soll über den Fortschritt berichtet werden). Etliche Aspekte wurden jedoch konkretisiert: u. a. der Verzicht, Urheberrechte geltend zu machen; die Möglichkeiten, über Nutzungslizenzen angemessene Entgelte für begrenzte Aufbereitungen (ggf. Anonymisierung und Fehlerbereinigung) definieren zu dürfen; die diskriminierungsfreie (Roh-)Datenbereitstellung; die Verwendung offener maschinenlesbarer Formate; die Bereitstellung dynamischer Datenbestände über offene Anwendungsprogrammierschnittstellen.

3.3 Umgang mit personenbezogenen Daten

3.3.1 Von Datenverarbeitung betroffene Personen und deren Rechte

Grundrechte

Einzelne natürliche Personen haben in Deutschland unterschiedliche unveräußerliche und dauerhaft einklagbare Grundrechte, die Verfassungsrang haben. Der Staat muss sie einerseits schützen, andererseits begrenzen sie die Staatsgewalt. Sie werden durch die Europäische Menschenrechtskonvention und auf nationaler Ebene durch das Grundgesetz[47] definiert. Im Kontext der Datenerhebung

45 Gesetz zur Förderung der elektronischen Verwaltung (E-Government-Gesetz – EGovG)
46 Gesetz zur Änderung des E-Government-Gesetzes und zur Einführung des Gesetzes für die Nutzung von Daten des öffentlichen Sektors
47 Grundgesetz für die Bundesrepublik Deutschland (GG)

und Verarbeitung vorrangig relevant sind das Recht auf Privatsphäre und das Recht auf informationelle Selbstbestimmung. Beide werden durch allgemeine Persönlichkeitsrechte begründet (Art. 2 Abs. 1 i.V. m. Art. 1 Abs. 1 GG). Das Recht auf Privatsphäre fußt auf der Annahme, dass ein privater, abgeschirmter Bereich nötig ist, um sich frei entfalten zu können und stellt diesen unter Schutz. Dieses Recht auf Privatsphäre wird in drei Bereichen explizit definiert und geschützt: in Bezug auf persönliche Informationen und Daten (informationelle Selbstbestimmung), Telekommunikation (Art. 10 GG) und Wohnung (Art. 13 GG). Das Recht auf informationelle Selbstbestimmung wurde vom Bundesverfassungsgericht 1983 im Volkszählungsurteil definiert und besagt, dass einzelne natürliche Personen grundsätzlich selbst über die Preisgabe und Verwendung ihrer personenbezogenen Daten bestimmen können (Schepers et al. 2015, S. 220). Grundrechte sind nicht schrankenlos. So wie Daten nicht nur ein Individuum, sondern auch soziale Realitäten einer Gemeinschaft abbilden, muss auch im Falle eines überwiegenden Allgemeininteresses die einzelne Person Einschränkungen ihrer Rechte hinnehmen, die gesetzlich zu definieren sind.

Die seit 2018 EU-weit gültige DSGVO zielt darauf ab, die unterschiedlichen Interessenlagen bezüglich des Schutzes von Grundrechten und der Verwendung personenbezogener oder -beziehbarer Daten abzuwägen. In der Verordnung werden Einzelpersonen gegenwärtig als Betroffene bezeichnet (Art. 4 DSGVO). Die Verarbeitung von personenbezogenen Daten ist nur rechtmäßig (Art. 6 DSGVO)

- mit freiwilliger Einwilligung der betroffenen Person zu bestimmten Zwecken, einschließlich der Erfüllung (vor)vertraglicher Maßnahmen, die auf Anfrage dieser Person erfolgen, sowie zum Schutz lebenswichtiger Interessen der betroffenen oder anderer Personen oder
- zur Wahrnehmung von Aufgaben im öffentlichen Interesse. Genannt werden u. a. Aufgaben im Gesundheits- und Sozialbereich (Art. 6 Abs. 1e DSGVO).[48] Dafür enthält die DSGVO *Öffnungsklauseln*, durch die Mitgliedsländer die Rechtmäßigkeit der Datenverarbeitung und eigenständige Verfahrensmodalitäten festlegen können; z. B. definieren
  - Sozialgesetzbücher (SGB), welche Daten zur Abrechnung sozialer Leistungen erhoben und für welche Zwecke von wem verarbeitet werden dürfen (Kap. 4.4.1),

48 Auch die innere und äußere Sicherheit und Gefahrenabwehr sind Aufgaben im öffentlichen Interesse (Art. 6, Abs. 1e DSGVO), die das informationelle Selbstbestimmungsrecht beschränken können. Diese Thematik wird im TAB-Projekt »Beobachtungstechnologien im Bereich der zivilen Sicherheit – Möglichkeiten und Herausforderungen« separat behandelt (www.tab-beim-bundestag.de/de/untersuchungen/u20900.html; 10.112021).

- diverse Registergesetze das öffentliche Interesse an speziellen Sachverhalten (vom Einwohnerregister bis zu Krebsregistern; Kap. 4.1.4) oder
- Statistikgesetze, welche personenbezogenen Daten Statistikämter erheben dürfen (z. B. Mikrozensus statt Volkszählung) und wer diese verarbeiten darf.

Damit ergibt sich bei der Verarbeitung personenbezogener Daten ein spezielles Verbotsprinzip mit Erlaubnisvorbehalt: Entweder hat die betroffene Person freiwillig zweckspezifisch und informiert zugestimmt (Standardmodell im privatwirtschaftlichen Bereich) oder die Erlaubnis ist gesetzlich definiert zur Wahrnehmung öffentlicher Aufgaben (Standardmodell im öffentlichen Bereich). Eine weitere Möglichkeit ist die Datenverarbeitung im Rahmen von sekundären Nutzungen, die mit einem ursprünglichen Zweck vereinbar sind z. B. zu Forschungszwecken; Kap. 3.3.4). In der Summe entsteht ein komplexes System von Erlaubnistatbeständen (Siemoneit 2018, S. 7).

Aus datenanalytischer Perspektive weitere besonders relevante Grundrechte sind das

- Recht auf Leben, körperliche Unversehrtheit und Freiheit (Art. 2 Abs. 2 GG), infolgedessen der Staat nur auf gesetzlicher Ebene in diese Rechte eingreifen darf und er verpflichtet ist, Grundrechte aktiv zu schützen;
- Gleichheitsgrundrecht, das Diskriminierung aufgrund bestimmter persönlicher Eigenschaften verbietet (national durch Art. 3 GG) und durch das Allgemeine Gleichbehandlungsgesetz[49] realisiert werden soll. Es ist ein für staatliches und privates Handeln geltendes Verbotsrecht, das auch die Entwicklung und den Einsatz von Algorithmen betrifft (ausführlicher z. B. in Orwat 2019; TAB 2020);
- Recht auf ungehinderten Informationszugang (national durch Art. 5 Abs. 1 GG), das durch unterschiedliche Informationszugangs- und -freiheitsgesetze realisiert werden soll, wodurch teilweise auch Daten und Informationen öffentlicher Einrichtungen u. a. für komplexe Datenanalysen zugänglich werden, sowie
- Recht auf Eigentum (Art. 14 GG), dessen Reichweite und Grenzen für materielle Sachen insbesondere durch das Privatrecht (§ 903 BGB) und für immaterielle Güter durch das Immaterialgüterrecht konkretisiert wird.

Grundrechterealisierung durch Datenschutzrechte

Auch wenn eine natürliche Person einer Datenerfassung und -verarbeitung freiwillig zugestimmt hat oder wenn dies auf gesetzlicher Grundlage erfolgte, bleiben deren Grundrechte erhalten. Die in der DSGVO definierten Rechte betroffe-

49 Allgemeines Gleichbehandlungsgesetz (AGG)

ner Personen sollen dies sichern.[50] Neben allgemeinen Rechten auf transparente Information, Kommunikation und Verfahrensmodalitäten (Art. 12 DSGVO) haben Betroffene bezüglich ihrer personenbezogenen Daten folgende Rechte gegenüber datenverarbeitenden Stellen:

- *Auskunft* zu Verfahrensmodalitäten, Datenerhebung und -verarbeitung (Art. 15 DSGVO), direkt gekoppelt mit Informationspflichten datenverarbeitender Stellen bei der Datenerhebung (Art. 13 und 14 DSGVO);
- *Berichtigung und Löschung* – dazu gehört auch das »Recht auf Vergessenwerden«, u. a., wenn die Daten für die definierten Zwecke nicht mehr nötig sind oder bei Widerruf der Einwilligung (Art. 16 und 17 DSGVO);
- *Einschränkung der Verarbeitung,* u. a., wenn die Richtigkeit geprüft werden muss, oder die Verarbeitung unrechtmäßig ist (Art. 18 DSGVO);
- *Datenübertragbarkeit,* d. h., einzelne Personen haben ein Recht, die sie betreffenden Daten selbst zu erhalten oder diese an eine andere verantwortliche Stelle zu übermitteln (Art. 20 DSGVO);
- *Widerspruch,* u. a. bei Datenverarbeitung für Aufgaben im öffentlichen Interesse sowie zu wissenschaftlichen und historischen Forschungszwecken (Art. 21 DSGVO).

Ähnlich wie es im Falle eines überwiegenden Allgemeininteresses möglich ist, Grundrechte gesetzlich zu beschränken, können auch die diese Grundrechte absichernden Datenschutzrechte in diesen Fällen gesetzlich beschränkt werden (Art. 23 DSGVO).

Rechte in Bezug auf automatisierte Datenverarbeitung

Die DSGVO gewährt jeder natürlichen Person ein Recht, keiner ausschließlich auf automatisierter Datenverarbeitung beruhenden Entscheidung unterworfen zu werden, die ihr gegenüber rechtliche Wirkung entfaltet (Art. 22 Abs. 1 DSGVO). Dazu gehört auch das *Profiling,* definiert als automatisierte Datenverarbeitung, mit der auf eine bestimmte Person bezogene Aspekte wie Arbeitsleistung, wirtschaftliche Lage, Gesundheit, persönliche Vorlieben, Interessen, Zuverlässigkeit, Verhalten, Aufenthaltsort oder Ortswechsel analysiert oder vorhergesagt werden (Art. 4 Ab. 4 DSGVO). Dies gilt nicht, wenn (Art. 22 Abs. 2 DSGVO):

50 Dieses Datenschutzrecht gilt in der EU nur für natürliche Personen. In einigen anderen Ländern gilt es auch für juristische Personen (z. B. Schweiz). In der EU haben juristische Personen ein Recht auf Geschäftsgeheimnisse, konkretisiert durch die Richtlinie 2016/94 über den Schutz vertraulichen Know-hows und vertraulicher Geschäftsinformationen (Geschäftsgeheimnisse) vor rechtswidrigem Erwerb sowie rechtswidriger Nutzung und das auf nationaler Ebene durch das Gesetz zum Schutz von Geschäftsgeheimnissen (GeschGehG) umgesetzt wird.

1. die betroffene Person ausdrücklich eingewilligt hat,
2. eine automatisierte Entscheidung im Rahmen eines Vertrages zwischen betroffener Person und datenverarbeitender Stelle erforderlich ist oder
3. es eine Rechtsvorschrift gibt, die entsprechende Vorgehensweisen zulässt (Öffnungsklausel für national eigenständige Regelungen zu Aufgaben im öffentlichen Interesse).

Das Bundesdatenschutzgesetz (BDSG) konkretisiert diesbezüglich die Zulässigkeit von

- Scoring und Bonitätsauskünften zum Schutz des Wirtschaftsverkehrs (§ 31 BDSG): Im Kontext von Vertragsverhältnissen können Wahrscheinlichkeitswerte über ein bestimmtes zukünftiges Verhalten einer natürlichen Person verwendet werden, wenn wissenschaftlich anerkannte mathematisch-statistische Verfahren dieses Verhalten nachweisbar erheblich vorhersagen können, dafür mehr als nur Adressdaten herangezogen werden und wenn bei Anschriftenmitbenutzung Betroffene unterrichtet wurden und dies dokumentiert ist.
- Automatisierter Entscheidungsfindung im Einzelfall einschließlich Profiling (§ 37 BDSG): im Rahmen der Leistungserbringung nach einem Versicherungsvertrag insbesondere bei der medizinischen Behandlung (Kap. 4.3).

Bezieht man die unterschiedlichen Rechte, die die DSGVO betroffenen Personen gewährt, auf den Data-Mining-Prozess, so wird hier eine gewisse Sonderstellung deutlich. Während Einwilligungs- und Widerspruchsrechte vor einer Datenverarbeitung wirken und Verarbeitungsbeschränkungs- sowie Löschungsrechte auch laufende Data-Mining-Prozesse beeinflussen können (Beschränkung der Datenerhebung und -verarbeitung), setzt dieses Recht voraus, dass Daten verarbeitet bzw. Data-Mining durchgeführt und Verfahren zum Scoring oder Profiling bereits entwickelt wurden (Beschränkung der operativen Anwendung von Profiling- oder Scoringverfahren).

Rechtsdurchsetzung

Die Datenschutzrechte betroffener Personen sollen zum einen durch definierte Grundprinzipien und Pflichten datenverarbeitender Stellen gewährleistet werden. Zum zweiten sollen Aufsichtsbehörden die Einhaltung überwachen und zum dritten haben betroffene Personen, die sich bezüglich ihrer datenschutzbezogenen Rechte verletzt sehen, ein Beschwerderecht bei einer Aufsichtsbehörde sowie ein Recht auf gerichtliche Klärung (Art. 77 und 79 DSGVO). Dafür ist seit 2018 auch ein Verbandsklagerecht vorgesehen (Art. 80 DSGVO). Ist aufgrund solcher Verstöße ein materieller oder immaterieller Schaden entstanden, hat die betroffene Person ein Recht auf Schadensersatz (Art. 82 DSGVO).

Jedes dieser Schutzelemente hat Stärken und Grenzen (ausführlicher z. B. Siemoneit 2018; Spindler et al. 2016). Trotz der etablierten Schutzelemente zur Rechtsdurchsetzung zeigen Meinungsumfragen regelmäßig, dass betroffene Personen skeptisch sind gegenüber vielfältigen datenverarbeitenden Stellen und bezweifeln, dass sie eine ausreichende Kontrolle über ihre Daten haben (stellvertretend z. B. Vodafone Institute for Society and Communications 2016). Die Verbesserung der Durchsetzung gewährter Grundrechte ist eine der gesellschaftlichen Herausforderungen im Kontext der Digitalisierung vielfältiger Lebensbereiche und der zunehmenden datenanalytischen Möglichkeiten. Diesbezügliche Fortschritte dürften der nach wie vor verbreiteten Skepsis u. a. gegenüber komplexen Datenanalysen in Deutschland entgegenwirken.

3.3.2 Grundsätze und Pflichten bei der Datenverarbeitung

Aufgrund des Marktortprinzips (Art. 3 und 27 DSGVO) gelten für alle datenverarbeitenden Stellen, die Dienste unter Verwendung personenbezogener Daten innerhalb der EU anbieten, egal wo sie ihren Sitz haben, folgende *Grundsätze* (Art. 5 ff. DSGVO):

- *Rechtmäßigkeit, Verarbeitung nach Treu und Glauben, Transparenz*: Personenbezogene Daten müssen auf rechtmäßige Weise und in einer für die betroffene Person nachvollziehbaren Weise verarbeitet werden;
- *Zweckbindung*: Die Verarbeitung ist nur für definierte, eindeutige und legitime Zwecke zulässig, jedoch wird eine Weiterverwendung für im öffentlichen Interesse liegenden Archivzwecken, für wissenschaftliche Forschungszwecke oder statistische Zwecke als nicht unvereinbar mit ursprünglichen Zwecken angesehen (privilegierte Datenweiterverwendung);[51]
- *Datenminimierung*: Die Verarbeitung soll dem Zweck angemessen und auf das notwendige Maß beschränkt bleiben.[52]
- *Richtigkeit*: Personenbezogene Daten müssen auf dem neuesten Stand sein, Fehler sind unverzüglich zu berichtigen bzw. fehlerhafte Daten zu löschen.
- *Speicherbegrenzung*: Daten, die eine Personenidentifizierung ermöglichen, dürfen nur solange gespeichert werden, wie es definierte Zwecke erfordern. Auch diesbezüglich ist eine Ausnahme für wissenschaftliche Forschungszwecke formuliert.
- *Integrität und Vertraulichkeit*: Datenverarbeitende Stellen müssen eine angemessene Sicherheit personenbezogener Daten gewährleisten und diese vor unbefugtem Zugriff und Verarbeitung schützen.

51 Diese Ausweitung der Zweckbindung gab es im früheren Datenschutzrecht nicht.
52 Datenminimierung wurde früher mit Datenvermeidung und -sparsamkeit assoziiert.

- *Rechenschaftspflicht*: Verantwortliche datenverarbeitende Stellen müssen die Einhaltung der Grundsätze nachweisen können.

Auf der Basis dieser handlungsleitenden Grundsätze werden explizite *Pflichten* für datenverarbeitende Stellen definiert (Kap. IV DSGVO):

- *Definition der Verantwortlichkeit und Dokumentation*: Eine verantwortliche Stelle und deren für den Datenschutz beauftragte Person sind zu benennen (ggf. auch eine bei beauftragten datenverarbeitenden Stellen). Verantwortliche und beauftragte Stellen müssen gemeinsam ein Verzeichnis über alle Verarbeitungstätigkeiten führen und der Aufsichtsbehörde auf Anfrage vorlegen (ausgenommen sind Unternehmen mit weniger als 250 Mitarbeitenden [Art. 30 DSGVO]).
- Sicherheit personenbezogener Daten bei der Verarbeitung durch angemessenes Datenschutzniveau:
  - Datenverarbeitende Stellen haben der jeweiligen Situation angemessene und geeignete technische und organisatorische Maßnahmen zum Schutz personenbezogener Daten zu ergreifen (Kap. 3.3.3), bei Daten besonderer Kategorie (Kasten 3.6) ist ein höheres Schutzniveau anzusetzen (Art. 32 DSGVO).
  - Es gibt umfangreiche Meldepflichten bei Datenschutzverletzungen gegenüber der Aufsichtsbehörde (Art. 33 DSGVO). Wenn eine Verletzung des Schutzes personenbezogener Daten mit einem voraussichtlich hohen Risiko für die persönlichen Rechte und Freiheiten natürlicher Personen einher geht, sind auch betroffene Personen zu benachrichtigen (Art. 34 DSGVO).
  - Insbesondere bei der Verwendung neuer Verarbeitungstechnologien müssen bei hohem Risiko der Rechteverletzung natürlicher Personen die datenschutzbezogenen Folgen durch solch eine Verarbeitung im Vorfeld abgeschätzt werden (Datenschutz-Folgenabschätzung [Art. 35 DSGVO]).

Kasten 3.6 Personenbezogene Daten besonderer Kategorie

Daten besonderer Kategorie bilden den Kernbereich der Persönlichkeit und der persönlichen Lebensführung ab. Die DSGVO nennt in einer nicht abschließenden Liste einerseits Daten zur eindeutigen Identifizierung natürlicher Personen (darunter biometrische und genetische Daten) und andererseits Daten, die als besonders diskriminierungssensibel gelten (darunter Gesundheitsdaten, Daten aus denen eine rassische und ethnische Herkunft, sexuelle Orientierung, religiöse oder weltanschauliche Überzeugungen, politische Meinungen oder Gewerkschaftszugehörigkeiten hervorgehen [Art. 9 Abs. 1 DSGVO]). Zum Kernbereich der persönlichen Lebensführung gehö-

ren zweifellos vielfältige weitere Kategorien von Gefühlen, Gedanken, Absichten, Gewohnheiten über Beziehungen zu anderen Personen bis zur wirtschaftlichen oder finanziellen Situation, die durch die Digitalisierung vielfältiger Lebensbereiche zunehmend anhand von Daten abgebildet werden können. Als Daten besonderer Kategorie werden diese bisher nicht explizit genannt.

Bezüglich der Verarbeitung gilt ein Verbotsprinzip mit Erlaubnisvorbehalt, jedoch sind diverse Ausnahmen gesetzlich definiert, darunter die Sicherstellung und Überwachung der öffentlichen Gesundheit oder die Gewährleistung der Gesundheits- und Sozialfürsorge. Bei diesen solle der Maßstab der Erforderlichkeit gelten. Kritiker/innen der derzeitigen Situation sind der Meinung, dass die Vielzahl der Ausnahmen auf eine weitgehende Freigabe der Verarbeitung hinauslaufen würde (Siemoneit 2018, S. 13).

Bei jeglicher Datenverarbeitung sind die Grundrechte betroffener Personen in besonderem Maße zu schützen. Oftmals bestimmen nicht nur die Dateninhalte an sich die Kategorisierung. Vielmehr hängt es von den jeweiligen Möglichkeiten der Datenverknüpfung und -verarbeitung ab, ob Daten und die daraus abgeleiteten Informationen den Kernbereich der persönlichen Lebensführung betreffen – nicht nur von der jeweils betroffenen Person, sondern auch von anderen Personen in deren Umfeld (z. B. bei genetischen oder sozialen Ähnlichkeiten).

Wenn dieser Kernbereich geschützt und erhalten werden soll, werden spezifische Betrachtungen von Datenverarbeitungsprozessen wichtiger, sowohl auf der Ebene einzelner Projekte mit einem Fokus auf unmittelbaren Folgen für jeweils betroffene Personen z. B. als Datenschutz-Folgenabschätzung, als auch auf übergeordneter Ebene mit einem gesellschaftlichen Fokus als Technikfolgenabschätzung.

Datenschutz-Folgenabschätzung

Datenverarbeitende Stellen können auch datenanalytische Projekte (darunter Data-Mining-Projekte) durchführen, bei denen ein hohes Risiko besteht, Grundrechte betroffener Personen zu verletzen. In diesem Fall müssen verantwortliche datenverarbeitende Stelle diese Risiken vorab genauer untersuchen und durch angemessene Maßnahmen möglichst vermeiden oder zumindest minimieren. Einige Länder verankerten bereits vor Jahren gewisse *Vorabkontrollen* oder *privacy impact assessments* in unterschiedlicher Ausprägung in ihren jeweiligen Datenschutzgesetzgebungen (Friedewald et al. 2017, S. 8 ff.). Durch die DSGVO gibt es erstmals verpflichtende Mindeststandards für eine Datenschutz-Folgenabschätzung (DSFA). Dazu gehören (Art. 35 und 36 DSGVO):

- eine systematische Beschreibung des geplanten Verarbeitungszwecks, der berechtigten Interessen der verantwortlichen Stelle sowie der Verarbeitungsvorgänge;
- eine zweckbezogene Bewertung sowohl der Notwendigkeit einer DSFA als auch der Risiken für die Rechte und Freiheiten betroffener Personen sowie
- geplante Risikominimierungsmaßnahmen (Garantien, Sicherheitsvorkehrungen).

Die verantwortliche datenverarbeitende Stelle führt eine solche Folgenabschätzung eigenverantwortlich durch. Nur wenn trotz Minimierungsmaßnahmen hohe Risiken in Bezug auf die Verletzung von persönlichen Grundrechten bestehen bleiben, muss vor der Datenverarbeitung die zuständige Aufsichtsbehörde konsultiert werden. Ein solches Risiko wird stets angenommen bei der umfangreichen Verarbeitung von Daten besonderer Kategorie (Kasten 3.6), beim Einsatz neuer datenanalytischer Verfahren (Kap. 2.3.2), bei umfassenden Bewertungen persönlicher Aspekte natürlicher Personen mittels automatisierter Analyseverfahren (u. a. Profiling), bei personenbezogenen Datenverarbeitungen im Kontext der systematischen Überwachung öffentlicher Räume (TAB 2022a) oder bei strafrechtlichen Verurteilungen.

Die DSGVO definiert lediglich Mindeststandards und lässt den Mitgliedsländern einen Umsetzungsspielraum bei der Übertragung in praktikable Anwendungen. Nationale Aufsichtsgremien sind z. B. aufgefordert, detailliertere Listen zu erstellen, aus denen hervorgeht, in welchen Situationen eine DSFA grundsätzlich durchzuführen bzw. nicht erforderlich ist. Derartige Positivlisten werden in Deutschland auf Bundes- und Landesebene erstellt und regelmäßig fortgeschrieben.[53] Darin werden u. a. unterschiedliche Verfahren zur Geolokalisierung sowie diverse medizinische und gesundheitssystemische datenverarbeitende Prozesse genannt (die Spanne reicht von Big-Data-Analysen und telemedizinischen Anwendungen über Anonymisierungsverfahren großer medizinischer Datenbestände zur Weiterverarbeitung zu anderen Zwecken oder KI-Anwendungen zur Beihilfefestsetzung durch Krankenkassen bis zu Bewertungsportalen für Ärzt/innen).

Unklar ist, inwiefern die Mitgliedstaaten an sich selbst bzw. ihre jeweiligen Ämter, die in der Regel eine vollständige Monopolstellung bei der Erfüllung öffentlicher Aufgaben haben und den Bürger/innen keine Wahlfreiheit bei Verwaltungsangelegenheiten lassen, die gleichen Anforderungen stellen, wie an privatwirtschaftlich agierende datenverarbeitende Stellen. Da diese öffentlichen Aufgaben gesetzlich zu definieren sind, kann eine DSFA bereits im Gesetzgebungsprozess erfolgen (Gesetzes-Datenschutz-Folgenabschätzung; Friedewald et al. 2017, S. 17). National wurde von dieser Option erstmalig im Rahmen des

53 www.bfdi.bund.de/DE/Fachthemen/Inhalte/Technik/Datenschutz-Folgenabschaetzungen.html (10.11.2021)

Digitale-Versorgungs-und-Pflege-Modernisierungs-Gesetzes Gebrauch gemacht und für unterschiedliche zulassungspflichtige Komponenten und Dienste der Telematikinfrastruktur des Gesundheitswesens (Kap. 4.1.3) eine zentrale DSFA durchgeführt, um zur Nutzung dieser Komponenten verpflichtete (medizinische) Einrichtungen von der Erstellung eine DSFA zu entlasten.[54]

Friedewald et al. (2017, S. 24 ff.) weisen darauf hin, dass bei einer DSFA eigentlich aus der Betroffenenperspektive prospektiv untersucht werden sollte, inwiefern ein datenanalytischer Prozess die Grundrechte Betroffener gefährdet. Wenn eine datenverarbeitende Stelle eine DSFA durchführt, gebe es stets einen Interessenkonflikt zwischen der eigenen Analyseabsicht bzw. dem eigenen Geschäftswillen und der Betroffenenperspektive. In einer solchen Situation würden datenverarbeitende Stellen tendenziell dazu neigen, die originäre Zweckbindung vorhandener Datenbestände zu überdehnen. Die Autoren sprechen sich dafür aus, eine DSFA von einer unabhängigen Instanz prüfen oder sogar ganz durchführen zu lassen und sie empfehlen, Betroffene bzw. deren Interessenvertretungen am Verfahren zu beteiligen. Bei datenverarbeitenden Stellen, die große Datenbestände aufbauen und/oder monopolartige Stellungen einnehmen sei dies besonders wichtig.

Für datenverarbeitende Stellen sind DSFA aufwendig, denn bisher gibt es noch keine klaren Vorgaben zur Durchführung. Erste Anwendungserfahrungen zeigen, dass vor allem für kleine und mittlere Unternehmen DSFA eine Herausforderung sind (Enquete-Kommission 2020, S. 67). Da die Länder gegenwärtig einen erheblichen Spielraum bei der Umsetzung haben, werden DSFA kaum von allen datenverarbeitenden Stellen in gleicher Qualität durchgeführt oder einheitliche Bewertungsmaßstäbe angewendet werden. Es wird eine Aufgabe der nächsten Jahre sein, die eingeführten Verfahren auf ihre Praktikabilität und Zielerfüllung zu prüfen und weiterzuentwickeln.

Grundsätzlich sind vorab durchzuführende produktbezogene Untersuchungen und Evaluationen Standardinstrumente des Vorsorgeprinzips. Aus der Vorsorgeperspektive ist die DSFA-Verankerung in der DSGVO ein wichtiger erster Schritt. Sie nimmt ausschließlich mögliche Grundrechteverletzungen bei einem datenanalytischen Prozess in den Blick. Zuverlässigkeits- Sicherheits-, oder Leistungsaspekte werden nicht untersucht und bewertet.

Eigenverantwortung, Haftung und Sanktionen

Mit der Einführung der DSGVO erhielten datenverarbeitenden Stellen mehr Eigenverantwortung, die Beweislast regelkonformen Handelns gegenüber Aufsichtsbehörden wurde umgekehrt. Statt jede Verarbeitung zu melden, müssen

54 Gesetz zur digitalen Modernisierung von Versorgung und Pflege (Digitale-Versorgung-und-Pflege-Modernisierungs-Gesetz – DVPMG; Datenschutz-Folgenabschätzung S. 1350 ff.

sie auf Anfrage nachweisen können, dass sie Daten verordnungskonform verarbeiten, wobei Unternehmen mit bis zu 250 Mitarbeitenden nicht zur Dokumentation verpflichtet sind (ausführlicher z. B. Siemoneit 2018).

Neben den definierten allgemeinen Grundsätzen und Pflichten und den Verfahrensvorgaben zur Realisierung enthält die DSGVO auch etliche Kann- oder Soll-Formulierungen, die einen datenschutzkonformen und verantwortungsvollen Umgang mit personenbezogenen Daten konkretisieren und befördern sollen:

- für einzelne Verarbeitungsbereiche können u. a. Verbände und Vereinigungen spezifische Verhaltensregeln erarbeiten (z. B. zur fairen und transparenten Verarbeitung, Pseudonymisierung oder Ausübung der Rechte betroffener Personen in der Medizin [Art. 34 DSGVO]);
- für datenverarbeitende Stellen sollen datenschutzspezifische Zertifizierungsverfahren sowie Datenschutzsiegel und -prüfzeichen eingeführt werden (Art. 42 DSGVO).

Unabhängige Aufsichtsbehörden sollen die Anwendung und Einhaltung der Regelungen überwachen (Art. 51 ff. DSGVO). Nichteinhaltungen können mit Geldbußen bis 20 Mio. Euro bzw. 4 % des Jahresumsatzes geahndet werden. Jenseits von Geldbußen können Mitgliedsländer weitere wirksame, verhältnismäßige und abschreckende Sanktionen erlassen (Art. 83 und 84 DSGVO). Diese Nichteinhaltung muss nicht zu einer tatsächlichen Schädigung betroffener Personen geführt haben. Ist durch eine Datenverarbeitung ein Schaden entstanden, haften die dafür verantwortlichen datenverarbeitenden Stellen (Art 82 DSGVO).

Privatwirtschaftliche und öffentliche datenverarbeitende Stellen

Privatwirtschaftlich agierende datenverarbeitende Stellen sollen sich im Rahmen des geltenden Rechts in Marktstrukturen möglichst frei entfalten können. Im Wettbewerb mit anderen Unternehmen können sie u. a. Data-Mining betreiben, Analysewerkzeuge und digitale Dienste entwickeln und anbieten. Es gilt das Privatrecht einschließlich DSGVO vollumfänglich. Grundlage jeglicher Verarbeitung personenbezogener Daten einschließlich Data-Mining ist aus rechtlicher Sicht die Einwilligung Betroffener, die laut DSGVO freiwillig, informiert und zweckgebunden sein soll. Unternehmen holen sie auf vertraglicher Basis durch Individualabreden[55] oder Allgemeine Geschäftsbedingungen (AGB) (§ 305 BGB) in breiter Form ein und informieren dazu anhand von Datenschutzerklärungen (ausführlicher z. B. Riehm 2018). Auf diese Weise erhalten sie regelmäßig umfangreiche Rechte zur Datenerhebung und -nutzung, die

55 Die Übergänge zwischen Individualabreden und AGB sind fließend. Es gibt unterschiedliche spezielle Individualabreden, z. B. medizinische Behandlungsverträge (Kap. 4.1.1).

vor allem dann vielfältige Weiterverwendungen einschließlich Data-Mining-Aktivitäten zulassen, wenn viele Personen die jeweiligen Digitalangebote eines Anbieters nutzen und dafür umfangreiche Datennutzungsrechte erteilen. AGB und Datenschutzerklärungen sind hochflexible rechtliche Instrumente, die sich digital schnell verändern lassen. Die Praxis zeigt, dass viele betroffene Personen nahezu blind in AGB einwilligen. Dazu kommt, dass marktdominierende Stellungen einzelner Anbieter oftmals kaum Alternativen zulassen. Die freiwillige, zweckgebundene und informierte Einwilligung gilt als eine der derzeitigen normativen Herausforderungen der Digitalisierung im Allgemeinen und komplexer Datenanalysen im Besonderen.

Öffentliche Einrichtungen als datenverarbeitende Stellen erfüllen gesetzlich definierte Aufgaben im öffentlichen Interesse, oftmals ohne Konkurrenz und Wettbewerbsstrukturen. Aufgaben und Rechte zur Datenerhebung und -verarbeitung werden gesetzlich definiert und nicht über Einwilligungen eingeholt. Die DSGVO enthält Öffnungsklauseln für den öffentlichen Aufgabenbereich. Dadurch haben die jeweiligen nationalen Normen des öffentlichen Rechts Vorrang. In Deutschland gelten bereichs- und aufgabenspezifische Regelungen (z. B. Sozialgesetzbücher [SGB]) vorrangig. Sie werden ggf. durch Bundes- bzw. Landesdatenschutzgesetze ergänzt [§ 1 Abs. 2 BDSG]). Der Gesetzgeber hat einerseits weitreichende Befugnisse, andererseits sind gesetzliche Anpassungen und Veränderungen aufwendig. Sofern dies notwendig und verhältnismäßig ist, können zur Erfüllung öffentlicher Aufgaben sowohl Grundrechte (im Kontext der Datenverarbeitung insbesondere das Recht auf Privatheit und informationeller Selbstbestimmung) als auch Datenschutzrechte betroffener Personen sowie Grundsätze und Pflichten bei der Datenverarbeitung beschränkt werden (z. B. Auskunftsrechte und Informationspflichten). Bei solchen Beschränkungen sollen jedoch die verfassungsgemäße demokratische Ordnung und der Wesensgehalt der Grundrechte und -freiheiten erhalten bleiben (Art. 23 DSGVO).[56] Dafür werden im Rahmen der Gesetzgebung die jeweiligen öffentlichen Interessen gegenüber den Interessen betroffener Personen abgewogen. Überwiegen erstere, können daraus öffentliche Aufgaben abgeleitet und die dafür notwendige Datenverarbeitung gesetzlich definiert werden. Entsprechende Gesetze müssen u. a. die jeweils zulässigen primären Verarbeitungszwecke, die dafür notwendigen Analysedaten, Verantwortlichkeiten, Speicherfristen, Grundrechterisiken sowie Maßnahmen gegen Missbrauch enthalten. Darüber hinaus können sie auch Möglichkeiten für sekundäre Weiterverarbeitungen eröffnen, u. a. um erhebliche Belange des Gemeinwohls zu wahren (§ 23 BDSG; Beispiele in Kap. 5).

56 Die Datenethikkommission empfiehlt Datenverarbeitungen, die diesen Wesensgehalt verletzen (z. B. Totalüberwachung, die Integrität der Persönlichkeit verletzende Profilbildungen) expliziter zu benennen und gegen derart ethisch unvertretbare Datennutzungen Maßnahmen zu ergreifen (DEK 2019, S. 19).

Da es für unterschiedliche öffentliche Aufgabenbereiche und deren Einrichtungen jeweils eigene spezifische Rechtsvorschriften gibt, die das BDSG lediglich ergänzt, unterschiedliche föderale Zuständigkeiten und Regulierungen zur Datenerhebung und -bereitstellung existieren sowie öffentliche und private Akteure und öffentliche und gebewerbliche Aktivitäten mitunter bereichsspezifisch verflochten sind, entstehen teilweise komplexe rechtliche Situationen. Diese Aufgabenteilungen und Verflechtungen wurden bereits bei der Erhebung und Bereitstellung von Geodaten sichtbar, im öffentlichen Gesundheitssystem sind sie noch vielschichtiger (Kap. 4 u. 5).

3.3.3 Grundrechteschützende Maßnahmen

Jede datenverarbeitende Stelle ist verpflichtet, personenbezogene Daten zu schützen und bei jeder Verarbeitung die Rechte der jeweiligen natürlichen Personen möglichst wenig zu gefährden. Daten besonderer Kategorie sind in besonderem Maße zu schützen. Welche Maßnahmen und Verfahren für welches Schutzniveau geeignet und angemessen sind, entscheiden sie situationsbezogen und ggf. im Rahmen einer Datenschutz-Folgenabschätzung, ggf. unter Einbeziehung von Aufsichtsbehörden. Diese Maßnahmen können an unterschiedlichen Stellen ansetzen: an den Daten, bei deren Weiterverwendung durch Dritte, an den Datenzugangsstrukturen, an den analytischen Vorgehensweisen und an den Analyseergebnissen. Alle Ansätze zielen darauf ab, bei jeglicher Datenverarbeitung betroffene Personen und deren Grundrechte zu schützen. Im Data-Mining-Kontext werden vor allem analysetechnische Ansätze teilweise unter dem Begriff »privacy-preserving data mining« subsummiert (z. B. Grosskreutz et al. 2010; Mendes/Vilela 2017; Schepers et al. 2015, S. 241 f.).

An den Daten ansetzende Schutzmaßnahmen

Direkt an den Daten ansetzende Schutzmaßnahmen sind regelmäßig ein aus rechtlicher Sicht zwingend notwendiges Element der Aufbereitung personenbezogener Daten, die der eigentlichen Datenanalyse im Data-Mining-Prozess vorgelagert ist (Abb. 2.1). Datenschützende und datenanalysierende Verfahren stehen in einem besonderen Spannungsfeld. Erstere sollen verhindern, dass Informationen über einzelne Personen nicht unautorisiert aus Datenbeständen herausdestilliert werden. Da jedoch die analytischen Möglichkeiten der Informationsextraktion kontinuierlich besser werden, müssen die diese Informationen schützenden Maßnahmen ebenfalls ausgebaut werden. Auch wenn die DSGVO bezüglich der direkt an den Daten ansetzenden Schutzmaßnahmen nur noch die Begriffe Verschlüsselung und Pseudonymisierung verwendet, ist im nationalen Recht nach wie vor der Begriff der Anonymisierung verankert. In der Praxis

werden Anonymisierungsverfahren massiv eingesetzt, um Datensätze und -bestände weiterverwenden zu können.

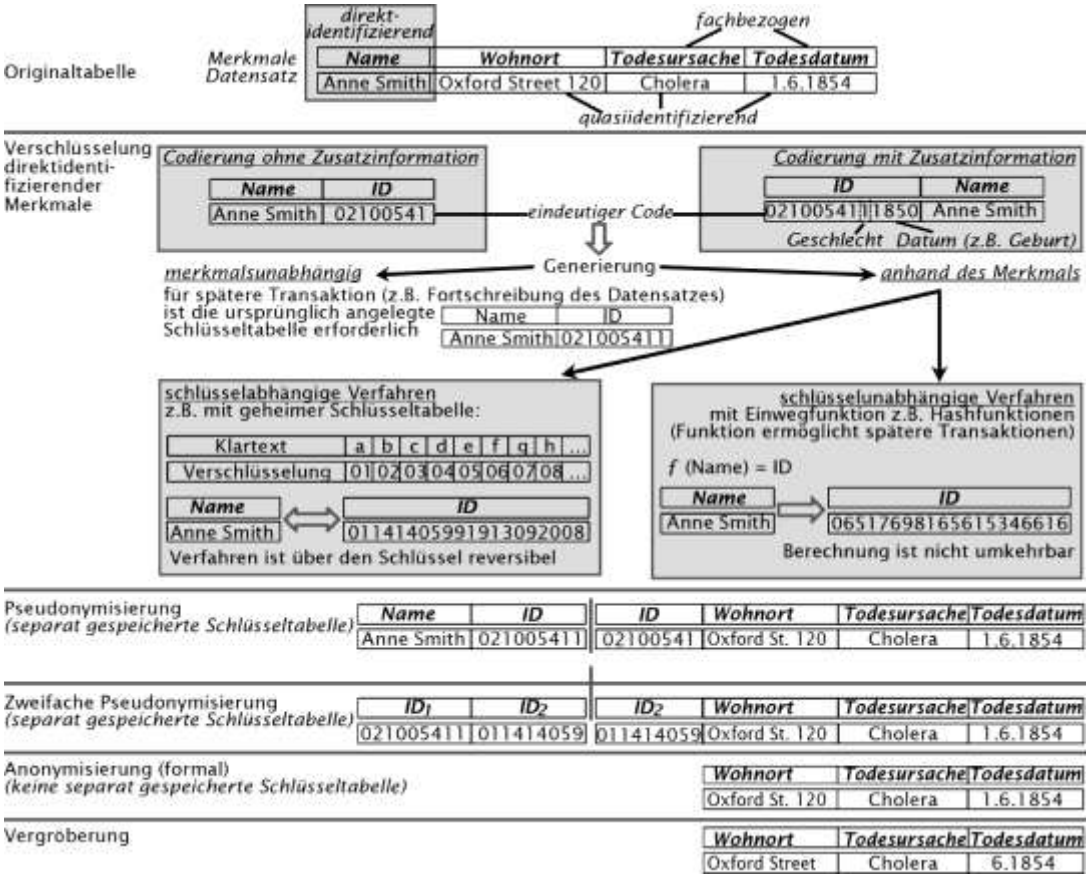
Bei der Pseudonymisierung und der Anonymisierung geht man davon aus, dass sich die Merkmale von Datensätzen drei unterschiedlichen Bereichen zuordnen lassen (Abb. 3.1 oben):

- *Direktidentifizierender Merkmalsbereich*: Merkmale bestimmen eine Person einzeln unmittelbar und eindeutig (z. B. Name plus Wohnort).
- *Quasiidentifizierender Merkmalsbereich*: Merkmale bestimmen eine Person mittelbar entweder in Kombination (z. B. Verknüpfung von Wohnort und Ereignis) und/oder durch die Verknüpfung mit zusätzlichen Daten (z. B. Verschlüsselung identifizierender Merkmale mit Zugang zur Schlüsseltabelle).
- *Fachbezogener Merkmalsbereich*: Merkmale repräsentieren eine fachliche Situation oder ein Ereignis, das auf mehrere Personen zutrifft.

Um die Grundrechte betroffener Personen als Datenobjekte zu schützen, können einzelne Merkmale, Merkmalsbereiche, ganze Datensätze oder auch -bestände mit unterschiedlichen Verfahren und Vorgehensweisen verschlüsselt oder anderweitig verändert werden. Ein Ansatz besteht darin, zunächst nur den direktidentifizierenden Merkmalsbereich (teilweise auch die quasiidentifizierenden) zu verschlüsseln (Pseudonymisierung) oder gänzlich aus den Datensätzen zu entfernen (Anonymisierung). Während bei der Pseudonymisierung der Personenbezug über die Verschlüsselungsverfahren grundsätzlich erhalten bleibt (Datenverarbeitung unterliegt weiterhin der Datenschutzgesetzgebung), hat die Anonymisierung den Anspruch, Personenbezüge von einem Datensatz gänzlich zu entfernen (diese Daten fallen danach nicht mehr unter die Datenschutzgesetzgebung). Ein zweiter Ansatz ist es, ganze Datensätze oder -bestände zu verschlüsseln, sodass nur diejenigen, die den Schlüssel kennen, das Verfahren rückgängig machen können. Ein dritter Ansatz verändert, vergröbert, verzerrt einzelne Merkmale, Merkmalsbereiche oder ganze Datenbestände. Oft werden die unterschiedlichen Ansätze miteinander kombiniert.

Für die *Pseudonymisierung* wird zunächst ein eindeutiger Code als Identifikator (ID) für jede Person erzeugt und dem Datensatz hinzugefügt (Abb. 3.1). Anschließend wird der direktidentifizierende Merkmalsbereich mit der ID-Kopie abgetrennt und in einer Schlüsseltabelle separat gespeichert. Es gibt unterschiedliche Codierungsverfahren. Alle zielen darauf ab, einen Code zu generieren, der jede Person eindeutig repräsentiert und die Re-identifizierung dieser Person durch Außenstehende, die keinen Zugang zu den Codierungsverfahren haben, verhindert. Mitunter werden zum eindeutigen Code (ebenfalls codierte) mitverschlüsselte Zusatzinformationen angehängt. Pseudonymisierte Datenbestände können mithilfe der Schlüsseltabellen oder Codierungsverfahren fortgeschrieben und ergänzt werden. Werden reversible Verfahren eingesetzt, kann der Vorgang, wie der Name schon sagt, rückgängig gemacht werden.

Abb. 3.1 Verschlüsselung, Pseudonymisierung, Anonymisierung und Vergröberung von Datensätzen



Eigene Darstellung

Die *Anonymisierung* hat den Anspruch, Personenbezüge irreversibel aus einem Datensatz zu entfernen. Identifizierende Merkmale werden vollständig abgetrennt und gelöscht, ohne dass Schlüsseltabellen angelegt oder gespeichert werden. Die Anonymisierung gewährleistet in der Regel zwar ein höheres Schutzniveau, lässt jedoch keine zeitliche Fortschreibung von fachbezogenen Merkmalen zu. Eine zuverlässige Anonymisierung wird mit der zunehmenden Verfügbarkeit vielfältiger Daten und deren Möglichkeiten zur Verknüpfung mehr und mehr in Frage gestellt.

Sowohl bei der Pseudonymisierung als auch bei der Anonymisierung werden auf der niedrigsten Personenschutzstufe nur direktidentifizierende Merkmale vom Datensatz abgetrennt (schwache oder auch formale Pseudonymisierung/Anonymisierung). Da man oftmals auch aus der Kombination mehrerer Merkmalsausprägungen auf eine Person schließen kann, kann auf der nächsten Stufe eine *Vergröberung* der quasiidentifizierenden Merkmale erfolgen (Abb. 3.1 unten). Die nächste Personenschutzstufe umfasst die Abtrennung auch

quasiidentifizierender Merkmale vom Datensatz (bei Pseudonymisierung mit, bei Anonymisierung ohne separate Speicherung in einer Schlüsseltabelle).

Anonymisierte einzelfallbezogene Datensätze können zusätzlich zu Fallgruppen zusammengeführt werden. Eine Standardgruppengröße gibt es nicht. Oft werden kleine Gruppen zusätzlich vergröbert, verschleiert oder unterdrückt (auch als k-Anonymisierung bezeichnet und als absolute Anonymisierung interpretiert). Der jeweilige Datentyp einzelner Merkmale bestimmt, ob und wenn ja, wie diese Zusammenführung erfolgen kann. Mittelwerte und Streuungsmaße können bei metrisch skalierten Daten (z. B. Alter, Gewicht) genutzt werden. Abstrakte Datentypen lassen sich mit derartigen Verfahren nicht zusammenfassen (es gibt weder Durchschnittsgenome noch Durchschnittsabbildungen oder -videos).

Teilweise können mittels Codierungs- und Klassifikationsverfahren abstrakte Datentypen in konkrete überführt und dann zusammengeführt werden. Vergröberungs- und Gruppierungsverfahren sind an sich nicht umkehrbar, sie verdichten die originären Daten, bei den resultierenden Datensätzen sinkt der Informationsgehalt. Alternativ können vor allem bei komplexen Datensätzen Verschleierungs- oder Verzerrungsverfahren eine Option sein (ausführlicher z. B. Mendes/Vilela 2017, S. 10565). Sie können auf unterschiedlich große Merkmalsbereiche und auch auf abstrakte Datentypen angewendet werden und in die Generierung synthetischer Datensätze und -bestände münden (ausführlicher z. B. DEK 2019, S. 132; Drechsler/Jentzsch 2018). Synthetische Daten haben in der Regel keinerlei Personenbezüge, da sie nicht in der realen Welt erhoben wurden, sondern aus Referenzdatenbeständen in großen Mengen abgeleitet werden. Die Datenethikkommission hält derartige Ansätze für vielversprechend (DEK 2019, S. 22), auch wenn sie eine Quelle für systematische Fehler sein können (Kasten 2.3). Sie empfiehlt eine Forschungsförderung, um entsprechende Verfahren weiterzuentwickeln.

Um auf einzelne Personen bezogene Datensätze (in der Statistik auch als Mikrodaten bezeichnet)[57] zeitlich fortschreiben zu können und dennoch eine Reidentifikation betroffener Personen möglichst auszuschließen, kommen unterschiedliche Vorgehensweisen zum Einsatz, die zwischen Pseudonymisierung und Anonymisierung anzusiedeln sind (Revermann/Sauter 2007, S. 157 ff.). Sie werden auch als *periodenübergreifende Pseudonymisierung* bezeichnet und sind für jegliche Längsschnittanalysen bzw. Zeitverlaufsuntersuchungen relevant (Data-Mining-Beispiel in Kap.5.5.3). Ein Ansatz ist eine mehrstufige Pseudonymisierung durch unterschiedliche voneinander getrennte Instanzen, die

57 In der amtlichen Statistik sind Mikro- oder Einzeldaten Daten(sätze), die Bezüge zu einzelnen natürlichen oder juristischen Personen haben und folglich der Geheimhaltung unterliegen. Das Gegenstück sind Makrodaten (aggregierte Datensätze, die sich auf unterschiedlich große Personengruppen beziehen und keiner Geheimhaltung unterliegen.

keinen Zugang zu den unterschiedlichen Schlüsseltabellen gewähren. Ein anderer Ansatz ist die Codierung mittels schlüsselunabhängiger Verfahren (Abb. 3.1). Diese generieren mit Einwegrechenanweisungen aus den identifizierenden Merkmalen nichtrückrechenbare eindeutige Codes, die diese Merkmale ersetzen. Dadurch gibt es keinerlei codifizierende Tabellen oder geheime Schlüssel, mit denen der Prozess rückgängig gemacht werden kann. Nur die geheimen Rechenanweisungen werden aufbewahrt, die bei einer Fortschreibung des Datensatzes aus den identifizierenden Merkmalen wieder den gleichen eindeutigen Code erzeugen (ausführlicher z. B. in Ertel 2012). Beim Einsatz schlüsselunabhängiger Codierungen ebenso wie bei Anonymisierungen können nachträglich keine Einwilligungen eingeholt oder fallbezogene Informationen ggf. zurückgegeben werden.

Die bisher genannten Verfahren verringern schrittweise den Informationsgehalt von Datensätzen (Schneider 2015, S. 251 f.). Eine Möglichkeit, um den Informationsgehalt vollumfänglich zu erhalten und die Daten dennoch vor unberechtigtem Zugriff zu schützen, ist deren *vollständige Verschlüsselung* beim Transport und der Speicherung. Dafür kommen nur reversible, also schlüsselabhängige Verfahren in Betracht. Es gibt symmetrische Verfahren (alle datenverarbeitenden Stellen nutzen den gleichen geheimen Schlüssel) und asymmetrische (Datensätze werden mit einem öffentlichen Schlüssel chiffriert und können nur mit einem privaten Schlüssel dechiffriert werden). Datensätze können damit nur bis zum nächsten Verarbeitungsknoten (Leitungsverschlüsselung) oder bis zum endgültigen Empfänger verschlüsselt werden (Ende-zu-Ende-Verschlüsselung). Bei der Übermittlung und Speicherung personenbezogener Daten besonderer Kategorie wird der Einsatz von Verschlüsselungstechniken explizit verlangt (Art. 32 DSGVO). Vollständig verschlüsselte Datensätze sind ohne Dechiffrierung einer Weiterverwendung und damit auch dem Data-Mining bisher weitestgehend entzogen. Jedoch wird an kryptografischen Verfahren geforscht, die trotz Verschlüsselung gewisse Analysen zulassen (homomorphe Verschlüsselungen). Sie befinden sich in frühen Entwicklungsstufen (Mendes/Vilela 2017, S. 10570). Welche der Schutzmaßnahmen im Einzelfall zum Einsatz kommen, wird je nach Verwendungszweck und der geforderten Reidentifizierungssicherheit entschieden.

Reidentifizierungssicherheit und deren Bewertung

Selbst wenn (quasi)identifizierende Merkmalsbereiche von Mikrodatensätzen verschlüsselt oder abgetrennt wurden, kann immer weniger ausgeschlossen werden, dass Einzelpersonen reidentifiziert werden können. Je umfangreicher Datensätze und -bestände sind, desto eher ist eine Reidentifizierung möglich (Kasten 3.7).

Kasten 3.7 Reidentifizierungsexperiment

Am US-amerikanischen MIT Media Lab wurde in einem Experiment eine pseudonymisierte Stichprobe aus den Verkehrs- bzw. Metadaten eines Kreditkartentransaktionsdatenbestand gezogen (Datum und Ort/Geschäft der Transaktionen von 1,1 Mio. Kunden in 3 Monaten) (Montjoye et al. 2015). Dann wurde anhand weniger Zeit- und Ortsdatenpaaren nach Pseudonymen gesucht, auf die dieses Transaktionsmuster zutraf. Das Experiments zeigte, dass mit den Zeit- und Ortsangaben von lediglich 4 Transaktionen 90 % der Kreditkartenkunden in der Stichprobe identifiziert werden konnten. Bei einer gröberen Auflösung (Vergröberung der Ort- und Zeitangaben zu 350 Geschäftsgruppen und Zeitintervallen von 15 Tagen) konnten mit 10 Merkmalspaaren 80 % aller Personen eindeutig identifiziert werden. Zudem konnten Frauen und Personen mit höheren Einkommen besser reidentifiziert werden als Männer und Personen mit mittleren oder niedrigeren Einkommen. Montjoye et al. geben auch zu bedenken, dass Personen durch Handynutzung oder Internetaktivitäten kontinuierlich raum- und zeitbezogene Datenspuren hinterlassen und Serviceanbieter sich an solchen Verkehrsdaten oftmals Rechte zur Weiterverwendung, teilweise auch zur Weitergabe sichern. Dies eröffnet ebenfalls Möglichkeiten der Reidentifizierung von anonymisierten Mikrodatensätzen, die in Folge erweitert und ergänzt sowie umfangreich analysiert werden könnten.

Die Bewertung der Sicherheit vor Reidentifizierung von Personen aus pseudonymisierten, anonymisierten oder vergröberten Mikrodatenbeständen ist eine kontinuierliche Herausforderung. Dafür werden mehrere Faktoren berücksichtigt (Revermann/Sauter 2007, S. 163 f.):

- der Umfang des jeweiligen Datenbestandes (je mehr Merkmale ein Datensatz enthält, desto wahrscheinlicher werden eindeutige Datenkonstellationen; Kasten 3.7),
- die Art und Häufigkeit der jeweiligen Dateninhalte (die Wahrscheinlichkeit auf eindeutige Datenkonstellationen steigt bei seltenen Merkmalsausprägungen) sowie
- Zahl, Art und Umfang verfügbarer Vergleichskollektive (mitunter auch als Zusatzwissen einer datenverarbeitenden Stelle bezeichnet).

Da sich diese Faktoren im Laufe der Zeit verändern, ist auch die Risikoabschätzung bezüglich der Personen-Reidentifizierung zeitabhängig. Was vor Jahren als sicher galt, kann durch größere Informationsverfügbarkeit und Zusatzwissen als risikoreicher bewertet werden. Wie das Reidentifizierungsexperiment verdeutlichte, müssten zum Schutz betroffener Personen immer größere Teile von Datensätzen und -beständen einer Analyse vorenthalten werden und dennoch

ließe sich eine Reidentifizierung einzelner Personen kaum ausschließen. Es scheint nicht mehr ausreichend, Pseudonymisierung und Anonymisierung aus der Sicht des Datenschutzes vorzuschreiben. Um Personen bei komplexen Datenanalysen einer Reidentifizierung zu schützen, werden vertrauensvolle Datenverarbeitungsstrukturen und Selbstverpflichtungen durch datenverarbeitende Stellen empfohlen (Art. 40 ff. DSGVO). Es gibt auch Forderungen, Reidentifizierungen explizit zu verbieten z. B. über Nutzungslizenzen (DEK 2019, S. 132).

Einwilligungsmanagement, Datentreuhänder und Infrastrukturen: kontrollierter Datenzugang vor allem für die Forschung

Neben den technischen gibt es diverse organisatorische Maßnahmen, die die Nutzung personenbeziehbarer Daten unter Einhaltung der Grundrechte betroffener Personen ermöglichen sollen. *Persönliche Datenmanagementsysteme* sollen betroffene Personen bei der Verwaltung ihrer Einwilligungen in die Datenweiterverwendung unterstützen. Sie sollen damit zunehmend differenziert und revidierbar festlegen können, wem sie welche Datenbestandteile zu welchen Zwecken zugänglich machen. Diese Selbstverwaltung informierter Einwilligungen ist jedoch voraussetzungsreich: Sie erfordert zum einen spezifische Kenntnisse bei betroffenen Personen und zum anderen transparente Darstellungen von Weiterverwendungsabsichten (ausführlicher z. B. in DEK 2019, S. 133 ff.). Entsprechende Ansätze stehen noch am Anfang der Entwicklung.

Datentreuhänder als vertrauenswürdige, kompetente, datenhaltende Stellen sind bereits seit Jahren etabliert. Relevant sind sie, wenn sensible Daten von vielen oder ganz bestimmten natürlichen oder juristischen Personen dauerhaft sicher gehalten und für unterschiedliche Untersuchungsfragen und komplexere Analysen nicht monopolisiert verarbeitet, sondern diversen datenverarbeitenden Stellen rechtssicher zugänglich gemacht werden sollen. Im juristischen Verständnis ist die Treuhänderschaft ein Rechtsverhältnis zwischen einem Treugebenden, der bestimmte Rechte, teilweise auch Pflichten an einem Treugut (hier Daten) hat, und einem Treuhänder, der diese Rechte im Außenverhältnis im eigenen oder auch in fremden Namen wahrnehmen darf (RfII 2020). Treugebende können sowohl betroffene Personen als auch datenverarbeitende Stellen sein, die Daten erhoben haben, und die diese für vielfältige Analysen bereitstellen wollen. Datentreuhandmodelle haben meist folgende organisatorische Elemente:

- *Datenannahmestellen* prüfen eingehende Datensätze, ggf. werden sie in Absprache mit den Datengebern korrigiert oder ergänzt;
- *Vertrauensstellen* realisieren nach den Korrekturen die Datenpseudonymisierung;

- *Registerstellen* verwalten, aktualisieren und erweitern umfangreiche pseudonymisierte Datenbestände, von denen sie Auszüge für Analysen ggf. bereitstellen;
- *Datenzugangskommissionen* prüfen Analyseanträge auf Legitimität, Rechtmäßigkeit und Einhaltung von Schutzstandards und genehmigen ggf. die Analyse (zunehmend werden sie zur transparenten Dokumentation verpflichtet) sowie
- *wissenschaftliche Serviceteams* unterstützen Analyst/innen bei der Realisierung ihrer Projekte, führen teilweise sicherheitskritische Analysen im Auftrag durch und prüfen die Personensensibilität von Ergebnissen vor der Weitergabe.

Man unterscheidet zwischen eigennützigen Datentreuhandmodellen, bei denen eine vertrauenswürdige juristische Person sowohl die datenhaltende als auch die datenanalysierende Stelle ist und fremdnützigen Datentreuhändern, die neutral zwischen datengebenden und datenanalysierenden Akteur/innen stehen und mögliche widerstrebende Interessen uneigennützig ausgleichen sollen. Um die Jahrtausendwende wurde mit den Forschungsdatenzentren der amtlichen Statistik ein fremdnütziges Datentreuhandmodell etabliert, das die im Rahmen der amtlichen Statistik erhobenen und dauerhaft zu haltenden Mikrodaten[58] für die wissenschaftliche Forschung bereitstellt (§ 16 BStatG).[59] Nutzungsberechtigt sind bisher ausschließlich Personen, die an Hochschulen, Universitäten und wissenschaftlichen Instituten beschäftigt sind (ab Gastwissenschaftsstatus). Antragstellende Einrichtungen werden beim Erstantrag geprüft. Eine Datennutzung darf nur für wissenschaftliche Forschungsprojekte beantragt werden (einschließlich Master-/Doktorarbeiten oder Aufträge von Ministerien). Die Datennutzung wird vertraglich vereinbart, ist zeitlich begrenzt und in der Regel kostenpflichtig. Nutzende werden zur Geheimhaltung verpflichtet, Reidentifizierungen explizit verboten. Der für eine Analyse notwendige Grad der Anonymisierung bestimmt die Form des Datenzugangs:

- *Formal anonymisierte Mikrodaten* (nur Namen und Anschriften sind abgetrennt, alle anderen Merkmale auch die regionale Zuordnung bleiben für die Analyse erhalten [§ 5a Abs. 3 Nr. 6 BStatG]) werden nur von Zentrumsmitarbeitenden analysiert. Nutzungsberechtigte liefern das Analyseprogramm. Die Ergebnisse werden auf Einzelfallsensibilität geprüft (d. h., vom Ergeb-

58 In der amtlichen Statistik gelten z. B. bei Häufigkeitsangaben bereits Gruppen mit 3 Objekten als anonymisiert (Moreau/Wolfsteiner 2017, S. 49). Beim Informationssystem Versorgungsdaten des Gesundheitswesens (Kap. 5.5.2) gilt eine Standardmindestfallzahl von 30 Versicherten, die in begründeten Einzelfällen auf 5 abgesenkt werden kann (DIMDI 2016, S. 36).

59 Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BstatG)

nis kann nicht auf ein einzelnes Datenobjekt geschlossen werden), Ergebnisse ggf. zusätzlich vergröbert oder unterdrückt. Nur geprüfte Ergebnisse werden zurückgesendet (kontrollierte Datenfernverarbeitung).

- *Faktisch anonymisierte Mikrodaten* (Datensätze können nur mit einem unverhältnismäßig hohen Aufwand an Zeit, Kosten und Arbeitskraft einer Person zugeordnet werden, eine Reidentifikation ist jedoch nicht gänzlich ausgeschlossen [§ 16 Abs. 6 BStatG]) können in einer geringen Vergröberung an Gastarbeitsplätzen der Forschungsdatenzentren analysiert werden (On-Site-Nutzung). In einer stärkeren Vergröberung werden sie als Scientific-Use-Files (SUF) an die jeweilige Forschungseinrichtung übertragen (Off-Site-Nutzung) (Bundesregierung 2016b, S. 29).
- Der breiten Öffentlichkeit werden nur *absolut anonymisierte Mikrodaten* auf Antrag als Public-Use-Files (PUF) gegen Entgelt oder in höherer Aggregation kostenlos zugänglich gemacht.

Laut Bundesregierung (2016b, S. 1, 29) soll mit dem Verfahren das Statistikgeheimnis gewahrt und ein Ausgleich im Spannungsverhältnis zwischen den Grundrechten auf informationeller Selbstbestimmung und Berufsfreiheit einerseits und der Wissenschafts- und Forschungsfreiheit (Art. 5 Abs. 3 GG) andererseits erreicht sowie nationales und europäisches Recht harmonisiert werden. Verfahren, mit denen betroffene Personen in eine Datenweiterverwendung zu Forschungszwecken einwilligen könnten, sind dabei weder vorgesehen noch möglich. Zum einen ist die Weiterverwendung zu Forschungszwecken durch die nationalen Statistikgesetze legitimiert. Zum anderen sind die Daten für die dauerhafte Speicherung zumindest zu pseudonymisieren, eine spätere Reidentifizierung für eine Kontaktaufnahme ist gar nicht zulässig. Forschungsdatenzentren befragen ihre Nutzergemeinschaft regelmäßig zu ihrer Zufriedenheit.[60] Positionen von nicht nutzungsberechtigten Stellen oder Betroffenen werden nicht berücksichtigt.

Protagonist/innen sehen in Datentreuhändern ein praktikables Brückenglied zwischen Datenschutz und Datennutzung. Mit ihnen können die Datenverwendung anhand definierter Kriterien gesteuert und analytische Prozesse überwacht werden. Treuhandverfahren seien eine Möglichkeit, Vertrauen und Akzeptanz in die Datenweiterverwendung u. a. zum Data-Mining zu erhöhen. Da Betroffene bisher weder der Datenerfassung für öffentliche Aufgaben noch der Datenweiterverwendung zu Forschungszwecken widersprechen können, bezeichnen Kritiker/innen solcherart treuhändische Weiterverwendung von Mikrodaten mitunter als paternalistische Fremdverwaltung, die de facto eine Entäußerung aller Entscheidungs- und Kontrollrechte betroffener Personen darstelle, und fordern eine Weiterentwicklung der Einwilligungsverfahren (Deutscher Ethikrat 2017, S. 181 ff.). Derzeit wird diskutiert, inwiefern persönliche

60 www.forschungsdatenzentrum.de/de/zufriedenheitsbefragung (10.11.2021)

Datenmanagementsysteme etablierte Treuhandverfahren ergänzen könnten, um das Grundrecht auf informationeller Selbstbestimmung in stärkerem Maße zu berücksichtigen. Expertenkommissionen empfehlen eine eingehende Befassung mit diesen (DEK 2019, S. 133; Deutscher Ethikrat 2017, S. 122; RfII 2020). Die Bundesregierung hat eine diesbezügliche Förderung zugesagt (Bundesregierung 2021a, S. 80).

Diese eingehende Befassung mit persönlichen Datenmanagementsystemen und die Weiterentwicklung bestehender Treuhandverfahren scheinen geboten, denn der Aufbau der Forschungsdatenzentren der statistischen Ämter war nur ein erster Schritt hin zum Auf- und Ausbau der Forschungsdateninfrastruktur (FDI) öffentlicher Einrichtungen. Sie ist ähnlich zur Geodateninfrastruktur als dezentrales Netzwerk diverser Forschungsdatenzentren öffentlicher Einrichtungen konzipiert. Neben den Mikrodaten der amtlichen Statistik vielfältige weitere Mikrodatenbestände aus anderen öffentlichen Aufgabenbereichen (vom Kraftfahrt-Bundesamt bis zu Rentenversicherungsdaten) sowie aus öffentlich finanzierten wissenschaftsgetragenen Erhebungen (vom sozioökonomischen Panel über Wahlumfragen bis zu Umfragen der Bundesregierung) nach definierten Standards gehalten und effizienter als vorher zu wissenschaftlichen Forschungszwecken zugänglich gemacht werden können (Stand November waren 41 Forschungsdatenzentren in der FDI akkreditiert bzw. hatten Gaststatus).[61]

Die Aktivitäten zum Auf- und Ausbau der FDI werden vom Rat für Sozial- und Wirtschaftsdaten (RatSWD) koordiniert. Er ist ein unabhängiges gewähltes Expertengremium, dessen Mitglieder ausschließlich aus datenbereitstellenden öffentlichen sowie aus datennutzenden wissenschaftlichen Institutionen stammen und von der Bundesregierung berufen werden. Interessenvertretungen von natürlichen oder juristischen Personen, die durch die Mikrodaten immer umfassender abgebildet werden können, oder Datenschutz- oder Datenaufsichtsinstitutionen sind nicht vertreten. Ratsinterne Arbeitsgruppen sichern vielfältige Aktivitäten (u. a. zur kontinuierlichen Qualitätssicherung und Verbesserung der Interoperabilität der Daten oder zur Outputkontrolle), thematisieren die strategische Weiterentwicklung und diskutieren u. a. forschungsethische oder gesellschaftliche Herausforderungen durch die immer vielfältigeren Möglichkeiten der Datenerhebung und -verknüpfung. 2009 wurden die Aktivitäten des RatSWD vom Wissenschaftsrat evaluiert (WR 2009): Die Erwartungen bezüglich der Verbesserung des Mikrodatenzugangs für wissenschaftliche Forschungseinrichtungen seien weit übertroffen worden. Der WR empfahl den RatSWD als Interessenvertretung der datenbereitstellenden und -nutzenden öffentlichen Institutionen in seinen Kernaufgaben zu stärken, um u. a. kontinuierlich weitere Datenquellen zu integrieren (u. a. Geodaten, administrative und privatwirtschaftliche Transaktionen, Daten zu Gesundheit, Psychologie und Bil-

61 www.ratswd.de/forschungsdaten/fdz (10.11.2021)

dung) und die disziplinäre Basis zu erweitern. Abgeleitete Aufgaben wie z. B. die Erstellung von Lernmodulen zur Forschungsethik und zum Datenschutz sollten abgegeben werden. Seit 2015 berichtet der RatSWD jährlich über die Entwicklung der FDI.[62]

Die unter Leitung des RatSWD aufgebaute FDI öffentlicher Einrichtungen soll perspektivisch mit der nationalen Forschungsdateninfrastruktur (NFDI) des Wissenschaftssystems vernetzt werden. Die NFDI soll primär zu Forschungszwecken erhobene Datenbestände aus vielen Fachdisziplinen nachhaltig sichern, zugänglich machen und international vernetzen. Erste NFDI-Aktivitätsschwerpunkte sind neben dem Aufbau der technischen Infrastrukturkomponenten, die Erschließung von Datenquellen, die Erarbeitung von Metadatenstandards (Interoperabilität) sowie die Stärkung datenverarbeitender Kompetenzen bei Forschenden, um eine intensivere Datennutzung zu ermöglichen (BMBF 2020). Die anvisierte Verknüpfung mit der forschungsgetriebenen European Open Science-Cloud und der sektorübergreifenden GAIA-X-Cloud (Kasten 2.2) sollen neben der Wissenschaft auch der Gesellschaft und der Wirtschaft einen besseren Zugang zu Forschungsdaten, Analysetools und -ergebnissen ermöglichen. Dadurch werden sich vielfältige Data-Mining-Möglichkeiten eröffnen. Beim Aufbau der Forschungsdateninfrastrukturen kann derzeit eine gewisse Aufbruchstimmung und eine erhebliche politischer Förderung beobachtet werden. Gesellschaftliche Herausforderungen beim Aufbau von Forschungsdateninfrastrukturen und der Intensivierung der Datennutzung zeichnen sich erst vage ab (ausführlicher z. B. Gehring 2018).

Analysekonzepte bei dezentraler Datenhaltung

Wenn Daten aus rechtlichen Gründen nur dezentral gehalten und nicht zusammengeführt werden dürfen und trotzdem Strukturen und Muster für eine Gruppe oder Gesamtheit ermittelt werden sollen, zeichnen sich Vorgehensweisen ab, die an die Möglichkeiten des parallelen Rechnens anknüpfen. Dabei verteilt eine Zentraleinheit Analysealgorithmen an teilnehmende datenhaltende Stellen, die damit ihre jeweiligen Teiledatensätze untersuchen, ermittelte Teilergebnisse auf Einzelfallsensibilität prüfen und nur solche Ergebnisse zurücksenden, die Dritten keine vertraulichen Informationen preisgeben. Eine Zentraleinheit führt ggf. die Teilergebnisse zusammen. Klassische Data-Mining-Verfahren z. B. zur Regelsuche, Gruppierung oder Klassifikation müssten dafür jedoch umformuliert werden, denn paralleles Rechnen ist jenseits einfachster Rechenoperationen schnell voraussetzungsreich und benötigt spezielle Protokolle und Algorithmen. Dezentrale Analyseansätze werden auch als sicheres verteiltes Data-Mining, oder allgemeiner als »secure multi-party computation« bezeichnet (Grosskreutz

62 www.ratswd.de/publikationen/taetigkeitsberichte (10.11.2021)

et al. 2010; Mendes/Vilela 2017; Schepers et al. 2015, S. 241 f.). Sie sollen automatisierte Datenanalysen bei gleichzeitiger Berücksichtigung von Datenschutzaspekten ermöglichen. Eine Grundvoraussetzung für derartige Ansätze sind semantisch und syntaktisch interoperable Datenbankstrukturen (Anwendungsbeispiel »Sentinel Initiative« der US-amerikanischen Arzneimittelzulassungsbehörde; Kap. 5.5.3).

3.3.4 Das Forschungsprivileg – ein Türöffner für Data-Mining

Seit 2018 gelten Weiterverwendungen personenbezogener Daten zu Zwecken, die im öffentlichen Interesse liegen, grundsätzlich als mit einem ursprünglichen Erhebungszweck vereinbar (*privilegierte Datenweiterverwendungen)*. Explizit genannt werden wissenschaftliche und historische Forschungszwecke sowie Archiv- und statistische Zwecke (Art. 5 Abs. 1 lit. b DSGVO). Diese Formulierungen öffnen Data-Mining als Knowledge Discovery in Databases in besonderem Maße die Tür. Die Freiheit zu forschen und wissenschaftlich zu arbeiten gehört in Deutschland und Europa zu den allgemeinen Grundrechten (Art. 5 GG; Art. 13 GRCh), die sowohl natürlichen als auch juristischen Personen, sowohl des öffentlichen als auch des privaten Rechts gewährt werden.

Wie bei der primären Datenverarbeitung müssen dabei technische und organisatorische Maßnahmen personenbezogene Daten schützen, um die Grundrechte betroffener Personen zu sichern (Kap. 3.3.1). Mit diesen Maßnahmen können ggf. sogar die in der DSGVO definierten Rechte betroffener Personen eingeschränkt werden, wenn sonst der Forschungszweck zumindest ernsthaft beeinträchtigt werden würde (Art. 89 DSGVO). Stärker als im eigentlichen Regelwerk der DSGVO wird in den erläuternden, aber rechtlich unverbindlicheren Erwägungsgründen der DSGVO ausgeführt, wie der Begriff *wissenschaftliche Forschung* verstanden werden soll:

- Sie solle im öffentlichen Interesse liegenden Zielen dienen und auf europäischen und nationalen Rechtsvorschriften beruhen (Erwägungsgrund 53 DSGVO).
- Sie sei mit der Einhaltung anerkannter ethischer Forschungsstandards verknüpft. Dazu gehöre auch ein datenbezogenes Einwilligungsmanagement. Da zum Zeitpunkt der Erhebung die Forschungszwecke oftmals noch nicht vollständig angegeben werden können, sollten betroffene Personen selektiv in Teile von Forschungsprojekten oder Forschungsbereiche einwilligen können (Erwägungsgrund 33 DSGVO).
- Auf dieser Basis solle der wissenschaftliche Forschungsbegriff weit ausgelegt werden, von der Grundlagenforschung bis zur angewandten Forschung reichen und auch technologische Entwicklungen und Demonstrationen sowie privat finanzierte Forschung einschließen (Erwägungsgrund 159 DSGVO).

Auch zur privilegierten Datenweiterverwendung zu Forschungszwecken ermöglichen Öffnungsklauseln nationale Konkretisierungen. In Deutschland gelten bezüglich der privilegierten Datenweiterverwendung zu Forschungszwecken diverse bereichsspezifische Regelungen des öffentlichen Rechts mit verteilter föderaler Zuständigkeit (für das Gesundheitssystem ausführlicher in Kapitel 5) vorrangig. Sie werden von Bundes- bzw. Landesdatenschutzgesetzen ergänzt. Das Bundesdatenschutzgesetz betont, dass die Verarbeitung personenbezogener Daten besonderer Kategorie für wissenschaftliche Forschungszwecke nur dann ohne Einwilligung zulässig ist, wenn die Verarbeitung erforderlich ist und die Forschungsinteressen den Interessen betroffener Personen an einem Ausschluss der Verarbeitung erheblich überwiegen (§ 27 Abs. 1 BDSG). Zu den Schutzmaßnahmen gehört eine möglichst baldige Anonymisierung (sofern berechtigte Interessen betroffener Person dem nicht entgegenstehen). Eine Reidentifizierung darf bei pseudonymisierten Daten nur erfolgen, wenn ein Forschungszweck dies erfordert. Auskünfte müssen nur erteilt werden, wenn der Aufwand dafür verhältnismäßig ist. Analysedaten dürfen nur mit Einwilligung betroffener Personen veröffentlicht werden (Ausnahme: Daten sind für die Darstellung der Forschungsergebnisse unerlässlich).

Etliche Formulierungen zur privilegierten Datenweiterverwendung zu Forschungszwecken sind auslegungswürdig (Siemoneit 2018). Unklar ist beispielsweise der Maßstab des öffentlichen Interesses in Bezug auf wissenschaftliche Forschung und deren Begrenzung. Inwiefern gehören Markt- und Meinungsforschung dazu? Betrifft das Einwilligungsmanagement auch öffentliche Aufgabenbereiche? Dürfen ausschließlich öffentliche Forschungseinrichtungen privilegiert werden? Welche ethischen Forschungsstandards sind heranzuziehen und bezüglich komplexer Datenanalysen bzw. Data-Mining relevant? Wie könnte die Einhaltung dieser Formulierungen kontrolliert werden, was passiert bei Verstößen? Die von der Bundesregierung eingesetzte Datenethikkommission (DEK) sieht in der Datennutzung für gemeinwohlorientierte Forschungszwecke (z. B. zur Verbesserung der Gesundheitsfürsorge) besondere Potenziale (DEK 2019, S. 20). Sie weist jedoch ebenfalls darauf hin, dass es bezüglich der Reichweite des Forschungsbegriffs und des Datenweiterverarbeitungsprivilegs im Zusammenhang mit der Entwicklung von Produkten Unsicherheiten gibt und empfiehlt diesbezügliche gesetzliche Klarstellungen.

Vorgehen in der medizinischen Forschung

Im medizinischen Bereich sind handlungsleitende Normen und Prinzipien der Schadensvermeidung, Vorsicht und Hilfe/Heilung (primum non nocere), Menschenwürde, Gerechtigkeit und Solidarität, der informierten Selbstbestimmung bzw. Patientenautonomie, der Geheimniswahrung und des Datenschutzes, der

wissenschaftlichen Güte, Verantwortung und Integrität der medizinischen Fachkräfte in der Professionsethik seit langem verankert. Sie gelten auch in der Forschung. Spezifische Gremien (Ethikkommissionen, Zulassungsbehörden, Datenzugangskommissionen) prüfen bei neuen Datenerhebungen und bei Anträgen auf Datenweiterverwendung, inwiefern ein Forschungsbedarf begründet sowie Forschungsinteressen mit den Interessen betroffener Personen abgewogen wurden (Kap. 4 u. 5). Diesbezügliche Prüfverfahren werden vor allem für größere Projekte u. a. wegen der verteilten Zuständigkeiten im föderalen Rechtssystem von Datenanalyst/innen vielfach als Barriere wahrgenommen. Perspektivisch sollen sie beschleunigt und vereinfacht werden (Bundesregierung 2021a, S. 19).

2017 prüfte der Deutsche Ethikrat die Anwendbarkeit und Passgenauigkeit (medizin)ethischer Prinzipien und Normen in Bezug auf komplexe Datenanalysen u. a. bei biomedizinischen Forschungsaktivitäten (Deutscher Ethikrat 2017). Er wies einerseits auf eine zunehmende technische Vernetzung medizinischer Einrichtungen und deren Datenbestände sowie ein grundsätzlich hohes Vertrauen betroffener Patient/innen in verantwortungsvolles Handeln datenverarbeitender Stellen hin. Andererseits hielt er die Kommunikationsmöglichkeiten zwischen den datenverarbeitenden Stellen und betroffenen Personen für mangelhaft, Einwilligungen in Datenweiterverwendungen können bisher kaum nachträglich eingeholt werden. Die sich kontinuierlich verbessernden technischen Möglichkeiten der Datenanalytik in Kombination mit fehlenden Einwilligungsmöglichkeiten seien eine der gegenwärtigen Herausforderungen bezüglich der Datenweiterverwendung zu Forschungszwecken im nationalen Gesundheitssystem (Deutscher Ethikrat 2017, S. 181 ff.). In Pilotprojekten werden neue Einwilligungsmodelle diskutiert und getestet (Kasten 3.8).[63]

Kasten 3.8 Einwilligungsmodelle

Bisher galten *breite Einwilligungen* (Broad Consent) in wissenschaftliche Forschungszwecke ohne Konkretisierung der Forschungsfragen zum Zeitpunkt der Erhebung in Kombination mit Datentreuhandstrukturen (Kap. 3.3.3) als praktikabelste Lösung (ausführlicher z. B. Deutscher Ethikrat 2017, S. 181 ff.; Rammos 2017). Aus ethischer Perspektive wird die de facto Entäußerung aller Entscheidungs- und Kontrollrechte betroffener Personen kritisiert.[64]

63 www.ceres.uni-koeln.de/forschung/projekte/leg2es/; www.medizininformatik-initiative.de/de/mustertext-zur-patienteneinwilligung (10.11.2021);

64 2020 wurde durch das Gesetz zum Schutz elektronischer Patientendaten in der Telematikinfrastruktur (Patientendaten-Schutz-Gesetz – PDSG) eine solche allgemeine Einwilligung in die Weiterverwendung von medizinischen Behandlungsdaten für wissenschaftliche Forschungszwecke im Rahmen der elektronischen Patientenakte ab 2023 gesetzlich verankert. Sie wird dort als Datenfreigabe oder -spende bezeichnet (§ 363 SGB V neu).

Eine Art Gegenmodell seien *dynamische Einwilligungen*, bei denen betroffene Personen selbst in einzelne Analyseanträge oder Teilprojekte informiert und zweckbestimmt einwilligen. Dieser Ansatz erfordert erhebliche Sachkenntnis, ist voraussetzungsreich und aufwendig. Die Gratwanderung zwischen ausreichend genauer und zu detaillierter Information (mit der Gefahr, dass nicht alles gelesen und verstanden wird) ist eine ständige Herausforderung (RatSWD 2020, S. 26). Auch sind direkte Kommunikationsmöglichkeiten zwischen betroffenen Personen und datenanalysierenden Stellen erforderlich, die es im nationalen Gesundheitssystem bisher nicht gibt.

Zwischen diesen beiden Ansätzen liegen *Kaskaden- oder Metaeinwilligungsmodelle*, bei denen betroffene Personen in einem ersten Schritt über die grundsätzliche Form (Broad Consent mit Datentreuhänder oder dynamische informierte Einwilligung) revidierbar entscheiden und danach grundsätzlich beide Wege eröffnet werden. Der Deutsche Ethikrat bezeichnete diese Form als derzeitigen Goldstandard (Deutscher Ethikrat 2017, S. 183 ff.). Sowohl dynamische als auch Kaskadeneinwilligungsmodelle benötigen persönliche Datenmanagementsysteme, mit denen betroffene ihre Einwilligungen definieren und verwalten können.

Andere Forschungsbereiche

Der von der Bundesregierung berufene unabhängige (Bei-)Rat für Sozial- und Wirtschaftsdaten (RatSWD) und dessen Arbeitsgruppe Forschungsethik empfehlen auch in anderen Wissenschaftsbereichen die Auseinandersetzung mit einer professionellen Ethik zu intensivieren und diese bereits in der Ausbildung, aber auch im Forschungshandeln stärker zu verankern (RatSWD 2017). Das der wissenschaftlichen Forschung grundsätzlich entgegengebrachte Vertrauen müsse durch verantwortungsvolles, kompetentes und transparentes Handeln und förderliche Strukturen gespiegelt und begründet werden. Dazu gehöre auch, bei jeder Daten(weiter)verwendung nicht nur die jeweiligen Forschungsinteressen in den Blick zu nehmen, sondern auch die Interessen betroffener Personen und deren Grundrechte u. a. auf informationelle Selbstbestimmung, Privatheit oder Gleichbehandlung zu achten und zu schützen. Dies erfordere auch einen gewissen Kulturwandel in der Scientific Community, der mit mehr Selbstreflexion, nicht mit mehr Bürokratie einhergehen solle. Dazu sollten bestehende forschungsethische Kodizes und Leitlinien (wie z. B. die zur guten wissenschaftlichen Praxis) zusammengeführt und neue datenanalytische Entwicklungen und Herausforderungen berücksichtigt werden. Zu beobachten sei bereits, dass vermehrt Ethikkommissionen auch jenseits der Medizin eingerichtet werden (ausführlicher z. B. Unger/Simon 2016). Diese könnten sowohl bei der Fortentwicklung und Verankerung der

wissenschaftlichen Forschungsethik einen Beitrag leisten, als auch bei der Interessenabwägung bei privilegierten Datenweiterverwendungen einen Beitrag leisten. Sie sollten als Datenzugangskommissionen konzeptionell in Dateninfrastrukturen eingebettet werden (RatSWD 2017, S. 24).

Die Deutsche Forschungsgemeinschaft (DFG) hat ethische Standards und Leitlinien guter wissenschaftlicher Praxis erarbeitet, deren Einhaltung Voraussetzung für deren Forschungsförderung ist (DFG 2013, 2019). Sie weist darauf hin, dass die Forschungsfreiheit untrennbar mit einer hohen Selbstverantwortung verbunden ist. Im Kontext der Datenerhebung und -verarbeitung verpflichtet sie ihre Mitglieder und Fördermittelempfänger gesetzliche und vertragliche Vorgaben einzuhalten sowie Vereinbarungen zu Nutzungsrechten an Daten und Ergebnissen zu treffen und zu dokumentieren. Bei einzelnen Forschungsvorhaben sollen deren Folgen gründlich abgeschätzt und ethische Aspekte beurteilt werden, ggf. sind Genehmigungen und Ethikvoten einzuholen. Um Forschungsergebnisse nachvollziehen und diskutieren zu können, sollen diese vollständig und nachvollziehbar beschrieben und möglichst öffentlich zugänglich gemacht werden, wobei die jeweiligen Autor/innen die Verantwortung für die Inhalte der Publikation tragen. Rohdaten, Methoden und Abläufe sowie Software (einschließlich Quellcodes für neue Analysesoftware) sollen offengelegt sowie fremde Vorarbeiten und Quellen vollständig ausgewiesen werden. Insbesondere bei der Erhebung von Daten und der Anwendung neuer Methoden sind Maßnahmen zur Qualitätssicherung von der Gerätekalibrierung bis zur Dokumentation einzuhalten. Bei der Datenaufbereitung und -bereitstellung sollen FAIR-Prinzipien eingehalten werden (Wilkinson et al. 2016): Metadatenfiles sollen Datensätze für Mensch und Maschine such- und auffindbar machen (Findable). Datensätze sollen offen, geschützte Daten über Treuhänder zugänglich gemacht werden (Accessible). Ausnahmen gibt es bei Patentanmeldungen. Standardisierte Begriffe und Formate sollen eingesetzt werden (Interoperability). Daten sollen bis zur Genese (eingesetzte Methoden und Geräte) rückverfolgbar sein, Nachnutzungsmöglichkeiten möglichst durch offene Nutzungslizenzen (Kasten 3.3) definiert werden (Reusability). Wissenschaftliche Einrichtungen sollen bis Mitte 2021 entsprechende technische Vorbereitungen treffen. Perspektivisch sollen die Datenrepositorien der Einrichtungen mit der nationalen Forschungsdateninfrastruktur, der European Open Science-Cloud (ein digitales Ökosystem, dessen Aufbau die Europäische Kommission von 2015 bis 2020 mit 600 Mio. Euro förderte) und der sektorübergreifenden GAIA-X-Cloud (Kasten 2.2) vernetzt werden.

3.3.5 Daten mit Bezug zu Personengruppen – (k)eine Sonderkategorie

Zwar ist die Unterscheidung zwischen Daten ohne und mit Personenbezug hilfreich, um sich den Rechtsraum des Datenumgangs grundsätzlich zu erschließen. Jedoch werden dazwischen liegende Graubereiche größer und bringen in Kombination mit einigen datenanalytischen Konzepten Kontroversen bezüglich der Zulässigkeitsgrenzen mit sich.

Die Zusammenführung von Mikrodaten zu Datensätzen, die sich auf kleine Personengruppen beziehen, ist ein Standardinstrument zur Anonymisierung personenbezogener Datensätze. Auch öffentliche Einrichtungen, die zunehmend zu Transparenz und Bereitstellung von Daten und Informationen verpflichtet werden, nutzen diese Instrumente u. a. zur Erstellung von »scientific« oder »public use files«, die datenverarbeitenden Stellen teils gegen Gebühr zugänglich gemacht werden. Solche sich auf Personengruppen beziehende Datensätze eröffnen seit Jahren lukrative Verwertungsmöglichkeiten: Eine ist das *Mikrotargeting* (TAB 2017b). Dafür werden zum einen datenbasierte Profile oder Features von Personengruppen erstellt. Über diese Features können auch Einzelpersonen als Datenobjekte Gruppen zugeordnet werden. Zum anderen haben datenverarbeitende Stellen eine Möglichkeit, Einzelpersonen gezielt anzusprechen (z. B. weil sie einen digitalen Service nutzen oder weil sie über die jeweiligen AGB von digitalen Kommunikationsplattformen oder anderen Serviceangeboten, Verlosungen oder Bonusprogrammen einer Weiterverwendung ihrer Adressdaten irgendwann einmal zugestimmt haben). Damit kann jegliche Informationsübermittlung von Aufklärung über Werbung bis zur politischen Botschaft nahezu individualisiert zielgruppenspezifisch direkt adressiert werden. In Deutschland hat z. B. die Deutsche Post Direkt GmbH das Informationssystem »DATAFACTORY« aufgebaut. Das Geschäftsmodell basiert auf zwei getrennten Datenbeständen:

- *»microdialog«* ist eine geografische Datenbank, deren Datenobjekte Mikrozellen mit durchschnittlich 6,6 Haushalten sind.[65] Diese Mikrozellen werden kontinuierlich angereichert, u. a. mit Daten zur Wohnsituation und zum -umfeld (u. a. vom Katasteramt), zur Fahrzeugausstattung (vom Datenzentrum des Kraftfahrt-Bundesamt regelmäßig als »public use files« gekauft), zur Kaufkraft, zum Bank- und Spendenverhalten (z. B. von Marktforschungsunternehmen) oder zu politischen Affinitäten (vom Meinungsforschungsinstitut dimap auf der Basis der amtlichen statistischen Wahler-

65 Nach derzeitigen Datenschutzauffassungen sind sechs Haushalte die kleinstmögliche Raumbezugseinheit, die nicht als personenbeziehbar gilt. Hat ein Gebäude mindestens sechs Haushalte, bildet es eine eigene Mikrozelle, bei kleineren Häusern erfolgt ein Zusammenschluss bis sechs Haushalte erreicht werden.

gebnisse auf Stimmbezirksebene). Für jede Mikrozelle werden für unterschiedliche werbe-/informationsrelevante Merkmale statistische Wahrscheinlichkeitswerte ermittelt. Nach Firmenangaben sind die 150 Einzelmerkmale von »microdialog »personenscharf« und dennoch datenschutzkonform.[66]

- Die *personenbezogene Adressdatei* enthält die postalischen Adressen von 46 Mio. Personen oder ca. 34 Mio. Haushalten in ca. 20 Mio. Häusern.[67] Laut Anbieter würde die Datei nahezu den gesamten Konsumentenmarkt abdecken. Sie ermöglicht es, Personen oder Haushalte direkt anzusprechen.

Laut Goldhammer/Wiegand (2017, S. 95 ff.) ist »microdialog« eines der vollständigsten und umsatzstärksten mikrogeografischen Systeme am deutschen Markt. Zielgruppen würden präzise charakterisiert (mit soziodemografischen, Konsum-, Struktur- und regionalen Daten sowie Branchen- und Lebensweltinformationen). Geschäftskunden erführen, welches Profil ihre Kunden hätten, wo es neue Potenziale gäbe, sodass Direktwerbung effizient die richtigen Empfänger erreichen könne. Das auf diese Datenbestände aufbauende Informationssystem »DATAFACTORY« mit seinen unterschiedlichen Funktionalitäten erlaubt Geschäftskunden u. a. spezifische Zielgruppen anhand von Merkmalen zu definieren. Entsprechende Mikrozellen werden selektiert und beispielsweise auf digitalen Landkarten visualisiert, die den Kunden für einen gewissen Zeitraum freigeschaltet werden. Auf Wunsch werden für diese definierten Mikrozellen auch separate Listen mit allen Adressdaten erstellt. Die Adressen werden nicht an die Geschäftskunden, sondern an einen kooperierenden Lettershop übermittelt, der das jeweilige Versandmaterial adressiert und verschickt und eine transparente Nutzung absichert (die für die Werbung verantwortliche Stelle wird benannt, Adressaten werden über ihre Werbewiderspruchsrechte informiert [Opt-out]). Weder Datenbestände noch das analytische Vorgehen werden offengelegt, (Geschäfts-)Kunden können nur das Front-End des Informationssystems nutzen. Die Aktivitäten der Deutschen Post Direkt GmbH unterliegen der Aufsicht der Bundesbeauftragten für Datenschutz, die bestätigte, dass das gewählte Verfahren im Einklang mit dem gültigen Datenschutzrecht steht. Solange das Ziel der Aktion eine gruppengenaue Information und Werbung sei, ohne dass daraus Entscheidungen mit rechtlicher Wirkung abgeleitet werden, gilt das Verfahren nicht als Profiling (Art. 22 DSGVO).

66 www.deutschepost.de/de/d/deutsche-post-direkt/microdialog.html (10.11.2021)

67 Diese Daten seien nicht aus der Postverkehrsdatei abgeleitet, sondern selbst zulässig erhoben worden. Für die Adresspflege wird jedoch u. a. die Anschriftenprüfung der Deutschen Post genutzt (https://ichsagmal.com/2018/04/03/verkauf-von-daten-fragen-an-deutschepostdhl-und-antworten-bitte-haben-sie-verstaendnis-dass-wir-darueber-hinaus-keine-o-toene-dazu-abgeben/; 10.11.2021).

Dennoch wurde das Vorgehen kontrovers auch im Bundestag diskutiert (Bundesregierung 2018a). Kritiker sehen die Grenze zum Missbrauch persönlicher Daten überschritten.[68] Sie sind der Meinung, ein Unternehmen, bei dem der Staat der größte singuläre Anteilseigner ist und das zudem einen faktischen Monopolstatus hat, würde personenbezogene Daten mit intransparenten Verarbeitungsverfahren an der Grenze zur Legalität kommerziell verwerten – sogar zu Wahlkampfzwecken und ohne dass die betroffenen Personen (Zielgruppen) zweckbestimmt und informiert zugestimmt hätten. Die Bundesregierung verweist auf das positive Votum der Bundesdatenschutzbeauftragen. Auch könne sie mittels ihrer Aktienanteile in Höhe von 21 % keinen Einfluss auf die operativen Entscheidungen der Deutschen Post AG nehmen (Bundesregierung 2018a, S. 8).

Dieses Beispiel soll veranschaulichen, wie sich Daten, auch solche, die im Rahmen öffentlicher Aufgaben ohne Einwilligung durch Betroffene erhoben wurden, unter Einhaltung existierender normativer Datenschutzvorgaben vielfältig anreichern lassen und dadurch ein breites Fundament für die Suche nach strukturellen Mustern bilden, anhand derer u. a. Kleingruppen gebildet und auch Personen zugeordnet und klassifiziert werden können, die selbst niemals in die Verwendung ihrer Daten zu solchen Zwecken eingewilligt haben. Auf dieser Datengrundlage können Informationssysteme entwickelt und Informationsdienste auch kommerziell vermarktet werden. Geschäftskunden erhalten keinen Zugang zur Datenbasis und müssen sich weder mit Belangen des Datenschutzes noch des analytischen Vorgehens beschäftigen. In den analytischen Maschinenraum des Informationssystems können sie nicht schauen – verkauft werden ihnen datenbasierte Informationen für ihre jeweiligen Aktivitäten.

Um abzusichern, dass derartige Prozesse und die dabei entstehenden datenanalytischen Werkzeuge dem Wohle der Menschen dienen, wird zunehmend gefordert, dass neben den Maßnahmen zur Sicherung des Datenschutzes auch die Folgen durch die Anwendung derartiger datenbasierter Werkzeuge und Informationssysteme für die Gesellschaft situativ in den Blick genommen werden (DEK 2019; Jaume-Palasí/Spielkamp 2017; Siemoneit 2018).

3.4 Umgang mit Data-Mining-Ergebnissen

Für diejenigen, die umfangreiche Datenbestände nutzen können, eröffnen sich unter Einhaltung von Datenschutzmaßnahmen vielfältige Analysemöglichkeiten. Im Rahmen von Data-Mining-Prozessen können sie Informationen generieren (Kap. 3.4.1) und datenanalytische Werkzeuge und digitale Anwendungen

68 www.faz.net/aktuell/wirtschaft/unternehmen/cdu-fdp-und-post-weisen-kritik-an-daten-geschaeften-zurueck-15522187.html; www.heise.de/tp/features/Datenmissbrauch-Deutsche-Post-und-CDU-nach-Facebook-und-Trump-4009611.html (10.11.2021)

entwickeln, die im Rahmen öffentlicher Aufgaben sowie ohne oder mit gewerblichen Absichten weiterverwendet werden können. Besonders Potenzial wird Ansätzen unterstellt, die mathematisch-statistische Modelle an Analysedaten anpassen oder Entscheidungsregeln ableiten, mit denen neue Situationen und Sachverhalte je nach definierter Untersuchungsaufgabe klassifiziert, gruppiert oder prognostiziert werden können – auch als algorithmische Assistenz- oder Entscheidungs(unterstützungs)systeme bezeichnet. Solche Systeme, vor allem wenn sie auf maschinellen Lernverfahren und neuronalen Netzen basieren, werden im Rahmen der derzeitigen KI-Euphorie massiv befördert. Bisherige Regulierungen greifen erst, wenn Data-Mining-Resultate zu (Software-)Produkten und Datenanalyst/innen zu Herstellenden werden (Kap. 3.4.2). Mit der Weiterentwicklung der diesbezüglichen rechtlichen Normen beschäftigen sich zahlreiche Expert/innen und Kommissionen (Kap. 3.4.3).

3.4.1 Informationen

Mit Data-Mining können vielfältige Informationen generiert werden, primär über Strukturen und Muster in Analysedatenbeständen, über Ähnlichkeiten, Zusammenhänge, Auffälligkeiten oder Unterschiede von Untersuchungsobjekten. Spezifische Fachkräfte und Datenanalyst/innen setzen sich mit diesen fachlich-inhaltlich auseinander, prüfen deren Aussagekraft, deren allgemeine Gültigkeit bzw. deren Generalisierbarkeit und verwenden sie im Rahmen der jeweiligen Tätigkeitsfelder weiter.

Der Umgang mit Informationen, die aus Data-Mining-Prozessen resultieren, wird bisher kaum spezifisch reguliert. Einerseits sollen Informationen im Rahmen von Forschungsaktivitäten möglichst frei genutzt und weiterverwendet werden können. Andererseits können sie im Rahmen von Innovationsprozessen und gewerblichen Aktivitäten als Geschäftsgeheimnis aufgefasst und exklusiv verwertet werden. Für öffentliche Aufgabenbereiche wird über unterschiedliche informationsbezogene Gesetze auf Bundes- und Landesebene der Umgang mit Informationen und Daten im Allgemeinen und diesbezügliche Offenlegungspflichten reguliert. Diese sind jedoch kaum Data-Mining-spezifisch.

Sofern durch Data-Mining-Prozesse generierte Informationen nicht nur organisationsintern, sondern für Dritte generiert und von diesen genutzt werden, wird diese Informationsgenerierung aus rechtlicher Sicht als (Dienst-)Leistung aufgefasst, die für Dritte im öffentlichen Aufgabenbereich auf gesetzlicher Grundlage und im privatwirtschaftlichen Bereich im Rahmen von Vertragsbeziehungen erbracht werden. Richtigkeits-, Sicherheits- und Haftungsfragen zur erbrachten Dienstleistung werden nur auf vertraglicher Ebene definiert, Verantwortlichkeiten des Leistungserbringers können weitgehend ausgeschlossen werden (eine Dienstleistung ist kein Produkt im Sinne des Produktrechts, ausführlicher z. B. Stock 2018, S. 89 ff.). Es besteht weitgehende Einigkeit, dass die

in der analogen Zeit entwickelte Regulierung bei digitalen Services, Informationsprodukten oder Informationsdiensten an ihre Grenzen kommt (siehe Mikrotargeting), Rechtsunsicherheiten bringt und die Rechte von Verbraucher/innen nur begrenzt schützen kann. Die derzeit auf europäischer Ebene in Abstimmung befindlichen Regularien zur künstlichen Intelligenz, über digitale Dienste und digitale Märkte sollen Grundrechte von Verbraucher/innen bei der Entwicklung und dem Einsatz von algorithmischen Systemen besser schützen (Kap. 3.4.3).

3.4.2 Algorithmen und Software

Data-Mining-Prozesse führen oftmals zu mathematisch-statistischen Modellen, Entscheidungsregeln, verallgemeinerbaren Rechenverfahren und Algorithmen, die zu digitalen Anwendungen oder Softwarebestandteilen entwickelt werden können, um sie organisationsintern einzusetzen oder auch Dritten bereitzustellen. Software wird bei jeglicher Form der Bereitstellung und Weitergabe auf dem Markt bzw. des Inverkehrbringens in Deutschland und Europa allgemein als Produkt aufgefasst (§ 2 ProdSG) [69] Mathematische Modelle, Regeln oder Algorithmen sind als Bestandteil von Software ebenfalls erfasst.

Das nationale Produktrecht hat eine komplexe Struktur. Produktsicherheit sowie Produzenten- und Produkthaftung sind miteinander verwoben, aber in unterschiedlichen Regularien gesetzlich verankert.[70] Die herstellende bzw. inverkehrbringende Stelle ist während der gesamten Lebensdauer eines Produktes für dessen Sicherheit verantwortlich. Werden erst im Laufe der Zeit Sicherheitslücken oder -fehler erkannt, müssen diese auch nachträglich beseitigt werden. Diese grundsätzliche Verantwortung kann in Deutschland und Europa vertraglich nicht umgangen werden.[71]

In einigen Anwendungsbereichen werden besondere Anforderungen sowohl an die Sicherheit als auch an die Leistungsfähigkeit von Hard- und Softwareprodukten gestellt (Bereiche mit erhöhter Kritikalität). In diesen Bereichen definieren eigenständige Rechtsnormen besondere Herstellerverpflichtungen

69 Gesetz über die Bereitstellung von Produkten auf dem Markt (Produktsicherheitsgesetz – ProdSG)

70 Das Produktsicherheitsgesetz wird ergänzt durch die Produzentenhaftung (§ 823 Abs. 1 BGB) und das Gesetz über die Haftung für fehlerhafte Produkte (Produkthaftungsgesetz – ProdHaftG).

71 Innerhalb des US-amerikanischen Rechtssystems können Softwarehersteller in Open-Source-Lizenzen bei kostenfreier Bereitstellung ihre Gewährleistungs- und Haftungspflichten gänzlich ausschließen. Im nationalen Rechtssystem können kostenlose Softwarebereitstellungen mittels Open-Source-Lizenz als Schenkung aufgefasst werden. Herstellende bzw. inverkehrbringende Stellen unterliegen der Produzentenhaftung und haften dann nur bei arglistigem Verschweigen von Mängeln bzw. grob fahrlässigem Handeln (§§ 523 u. 524 BGB).

und komplexere Leistungsnachweis- und Sicherheitsarchitekturen, um die Sicherheit beim Einsatz entsprechender Hard- und Softwareprodukte zu gewährleisten. Die Einhaltung dieser definierten Anforderungen wird geprüft, zertifiziert und mit einer CE-Kennzeichnung bescheinigt. Oftmals sind definierte Prüfinstanzen daran beteiligt. Aus der Data-Mining-Perspektive relevant sind bereits heute die Sicherheitsarchitekturen u. a. für Medizinprodukte und für IT-Systeme, die als Elemente kritischer Infrastrukturen gelten:

- Software, die zur Diagnose und Behandlung von Krankheiten eingesetzt werden soll, gilt als Medizinprodukt und wird je nach potenzieller Gesundheitsgefährdung einer von vier Risikoklassen zugeordnet. Hersteller müssen im Rahmen der Produktentwicklung sowohl Sicherheits- als auch Leistungsnachweise erbringen und ggf. Prüfinstanzen hinzuziehen. Für die breite Anwendung jenseits der Entwicklung ist eine Zertifizierung der Software und die Etablierung eines gestuften Risikomanagementsystems erforderlich (ausführlich in Kap. 4.2).
- Betreiber kritischer Infrastrukturelemente[72] müssen ihre IT-Komponenten einschließlich Datenrepositorien, Informationsdienste, Netze, Portale und Webangebote gegen unbefugte Zugriffe (Cyberangriffe) technisch absichern, zertifizieren und zweijährlich prüfen, ggf. das Bundesamt für Sicherheit in der Informationstechnik (BSI) als Prüfinstanz hinzuziehen und dem BSI erhebliche Störungen melden (Kasten 3.1).

Prüfinstanzen spielen in allen Sicherheitsarchitekturen risikobehafteter Produkte eine wichtige Rolle. Sie sollen u. a. den Stand der Wissenschaft und Technik bezüglich Risikominimierung und möglicher Produktfehler definieren, Standards für produktspezifische Bewertungsverfahren erarbeiten und deren Einhaltung prüfen, die jeweiligen Produkte zertifizieren/zulassen sowie das spezifische Marktgeschehen überwachen. Bei Medizinprodukten wird einerseits die fachlich-inhaltliche Leistung und andererseits die informationstechnische Sicherheit geprüft und zertifiziert. Dafür können Prüfinstanzen unter Wahrung von Geschäftsgeheimnissen auch Einblick in Analysedaten, methodische Vorgehensweisen und Unterlagen verlangen (ausführlich in Kap. 4.2). Für derartige Prüf- und Kontrollaktivitäten sind umfangreiche Kenntnisse und Ressourcen erforderlich (FDP 2019)

72 Kritische Infrastrukturen (KRITIS) sind für das Gemeinwesen von großer Bedeutung und müssen in besonderem Maße prospektiv vor möglichen Gefährdungen gesichert werden. Die Bewertung von IT-Komponenten als kritische Infrastrukturkomponenten ist in ständiger Entwicklung. Das nationale IT-Sicherheitsgesetz und inzwischen neun bereichsspezifische BSI-KRITIS-Verordnungen (Verordnung zur Bestimmung Kritischer Infrastrukturen nach dem BSI-Gesetz [BSI-Kritisverordnung – BSI-KritisV]) definieren spezifische Vorgehensweisen und Verantwortlichkeiten.

Die Verantwortung für die Produktsicherheit wird ergänzt durch eine eigengesetzlich definierte deliktische *Haftung* im Schadensfall. Bisher haften herstellende oder inverkehrbringende Stellen nur, wenn durch Softwarefehler größere Personen- oder Sachschäden (über 500 Euro) entstehen, nicht aber bei immateriellen oder digitalen Schäden (u. a. Datenveränderungen oder -verluste, Fehlinterpretationen und -bewertungen, Sperrungen oder Ausschlüsse von digitalen Aktivitäten). Ein Fehler liegt vor, wenn die Software nicht die Leistung und Sicherheit bietet, die berechtigterweise erwartet werden kann. Herstellende bzw. inverkehrbringende Stellen haften bei erkennbaren Fehlern, nicht aber bei Entwicklungsrisiken, wenn ein Fehler nach dem Stand der Wissenschaft und Technik beim Inverkehrbringen nicht erkannt werden konnte (§ 1 Abs. 2 Ziff. 5 ProdHaftG). Sie können bereits Entwicklungsversionen bereitstellen (Beta-Versionen) und Fehler durch Updates nachträglich beheben. Kontinuierliche Überarbeitungen und neue Versionen machen Software zu sich kontinuierlich weiterentwickelnden Produkten. Die Entwicklungsdynamik wird erhöht, wenn Software lernende Systembestandteile hat, deren Leistungsfähigkeit sich verändern kann (Kap. 2.3.2). Die Europäische Kommission beschäftigt sich im Rahmen ihrer KI-Strategie auch mit der Weiterentwicklung des Haftungsrechts. Dazu werden derzeit öffentliche Konsultationen durchgeführt (bis Januar 2022).[73]

Bei Data-Mining-Prozessen wird Analysesoftware von datenverarbeitenden Stellen eingesetzt, um Strukturen oder Muster in Daten zu erkennen. Auch wenn die mit der Software ausgeführten Rechenprozesse fehlerhaft sind, dürften sie kaum unmittelbare Schäden verursachen, die von der Produkthaftung erfasst werden – die Software trainiert und parametrisiert zunächst nur allgemeine mathematisch-statistische Modelle oder ermittelt Entscheidungsregeln. Tatsächliche Schäden entstehen erst, wenn diese Modelle und Regeln in neuen Situationen eingesetzt werden, um Entscheidungen zu unterstützen oder sogar automatisiert zu treffen. Dazu müssen diese nicht unbedingt in den Verkehr gebracht werden. Nur intern genutzte Softwareelemente, die nicht in den Verkehr gebracht werden und z. B. nur Informationsanfragen bearbeiten (z. B. Suchmaschinen, Wetterdienste), werden in der Regel vom Produktrecht nicht erfasst. Zudem können sich datenanalytische Werkzeuge und Software in global angelegten Cloudstrukturen europäischen und nationalen Rechtsstrukturen teilweise entziehen.

3.4.3 Rechtsunsicherheiten und Entwicklungsinitiativen

Seit Jahren wird auf gewisse Rechtsunsicherheiten im Kontext komplexer Datenanalysen hingewiesen und die diesbezügliche Zukunftsfähigkeit unterschiedlicher Rechtsbereiche diskutiert. Bei der Verarbeitung personenbeziehbarer Daten ergeben sich Unsicherheiten aus der DSGVO (Kap. 3.3.2): Sie verlangt

73 https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12979-Civil-liability-adapting-liability-rules-to-the-digital-age-and-artificial-intelligence/public-consultation_de (10.11.2021)

Transparenz und Nachvollziehbarkeit bei jeglicher Verarbeitung, was aus analytischer Sicht Unsicherheiten bezüglich der Anforderungen an das Design von Berechnungsverfahren, Algorithmen oder Software mit sich bringt. Auch gewährt sie natürlichen Personen ein allgemeines Recht, keiner ausschließlich auf automatisierter Datenverarbeitung beruhenden Entscheidung unterworfen zu werden, die ihnen gegenüber rechtliche Wirkung entfaltet (Kap. 3.3.1). Dieses Recht Betroffener wird rechtlich bisher weder durch Pflichten datenanalysierender Stellen gespiegelt, noch in Ge- oder Verbote zur Entwicklung und Anwendung solcherart automatisierter Entscheidungssysteme übersetzt. Zudem sind die Vorgaben der DSGVO in vielfältigen Anwendungsbereichen nicht bindend, weil für die Entwicklung solcher Systeme anonymisierte oder per se nichtpersonenbezogene Daten verwendet wurden und entstehende Regeln und Klassifikationsverfahren keinerlei Personenbezüge haben. Nicht die Entwicklung, sondern die Anwendung solcher Regeln und Verfahren kann (Grund-)Rechte und Freiheiten auf individueller Ebene oder die freiheitliche Grundordnung auf überindividueller Ebene verletzen. Dies betrifft unterschiedliche Datenbereiche.

Bernsdorf et al. (2015, S. 213 f.) haben mit Vertretern der GeoIT-Branche diskutiert, inwiefern bei komplexen Datenanalysen unter Verwendung von Geodaten ausreichende Rechtssicherheit bezüglich der zulässigen Verarbeitung besteht oder ob ein eigener Rechtsrahmen für Data-Mining-Aktivitäten nötig sei. Es wurde darauf hingewiesen, dass es zwar sehr detaillierte gesetzliche Vorgaben zur Erhebung und strukturierten Bereitstellung von (Geo-)Daten vor allem im Rahmen öffentlicher Aufgaben gibt, die Möglichkeiten und Grenzen der Analytik würden bisher jedoch kaum normativ konkretisiert werden (Kap. 3.2). Grundsätzlich können über die Geodateninfrastrukturen und über Nutzungslizenzen (Grund-)Rechte von Datengebenden geachtet sowie Möglichkeiten und Grenzen der Datenverwendung situativ vereinbart werden. Es fehlen jedoch ein übergeordneter Rechtsrahmen sowie analysespezifische Leitlinien für eine verantwortungsvolle Datenanalytik und Entwicklung zunehmend automatisierter Entscheidungs(unterstützungs)systeme. Vor allem bei hochaufgelösten Geodaten würden datenanalysierende Stellen die Situation als rechtlich kompliziert wahrnehmen, oftmals gäbe es einen Interpretationsspielraum bei der Bewertung der Kritikalität von Daten, von Analysen und von möglichen Ergebnissen. Unsicherheiten bestünden oftmals bei der Grenzziehung: Welche Klassifikation von (Geo-)Objekten, welche Analyse von Mobilitätsdaten sind noch zulässig, wann werden Analysen bedenklich oder (Grund-)Rechte gefährdet?

In etlichen Berichten, Weißbüchern und Strategien werden sowohl die besonderen Potenziale vor allem datenbasiert lernender Systeme zur Unterstützung von Entscheidungen betont (auch als algorithmische oder KI-Systeme bezeichnet), aber auch auf damit verbundene Unsicherheiten im derzeitigen Rechtssystem hingewiesen und Vorschläge zur Weiterentwicklung der derzeitigen rechtlichen Strukturen gemacht (u. a. DEK 2019; EK 2020d; Enquete-

Kommission 2020). Die Datenethikkommission spricht sich dafür aus, die lange Zeit dominierenden Datenschutz- und Datenzugangsperspektiven um *Algorithmenperspektiven* zu ergänzen (DEK 2019, S. 77). Damit könne die Dynamik datenbasiert lernender Systeme zur Entscheidungsfindung mit ihren Wirkungen und Folgen auf und für einzelne Personen und die Gesellschaft, die teilweise gar keine Datengeber waren, aber dennoch (und zunehmend) von deren Einsatz und deren Ergebnissen betroffen sein werden, besser adressiert werden.

Insbesondere bei KI-Systemen spricht sie sich für eine intensivere gesellschaftliche Auseinandersetzung sowie für die Weiterentwicklung des Rechtsrahmens insbesondere des Produktrechts aus. Auch andere Stimmen halten das gegenwärtige Produktrecht diesbezüglich für ungeeignet. In digitalen Kontexten, in denen Daten zunehmend ubiquitär erhoben, gehalten und verarbeitet werden und die Entwicklung von Systemen zur Entscheidungsunterstützung kaum reguliert sei, könnten Verantwortlichkeiten und Haftungsfragen nicht eindeutig zugewiesen und adäquat beantwortet werden (Rott 2018). Es gebe erheblichen Reformbedarf u. a. zum geschützten Rechtsgut oder zur Beweislast bei Haftungsfragen. Da das jetzige Verfahren geschädigte Personen typischerweise von der Verfolgung ihrer Ansprüche abhält, könne es weder softwareherstellenden noch -anwendenden Personen gegenüber Wirkung entfalten. Es gibt unterschiedliche Meinungen, ob primär die Regelungen zur Produktsicherheit weiterentwickelt und/oder Haftungsfragen spezifischer reguliert werden sollten. Zudem müsse im europäischen Binnenmarkt auch eine EU-weite Lösung gefunden werden. Erste Vorschläge zur Weiterentwicklung des Rechtsrahmens in Bezug auf KI-Systeme wurden auf europäischer und auf nationaler Ebene erarbeitet (DEK 2019; EK 2020d, 2021).

Die Europäische Kommission (EK 2020d) spricht sich dafür aus, eine möglichst hohe Produktsicherheit von KI-Systemen regulativ zu gewährleisten. Sie plädiert für eine Zweiteilung des künftigen Rechtsrahmens, der auf einem risikobasierten Bewertungsansatz aufbaut. Bei KI-Systemen mit geringem Risiko hält sie die europäischen Rechtsvorschriften zum Schutz von Grundrechten sowie zur Produktsicherheit und -haftung grundsätzlich für ausreichend, befürwortet jedoch freiwillige Kennzeichnungen und Gütesiegel im Rahmen der Selbstverwaltung. Erste Konzepte für derartige Gütesiegel oder Ethiklabel, die ethische Prinzipien und Grundsätze datenverarbeitender Stellen u. a. zur Transparenz, Verantwortlichkeit, Privatsphäre, Gerechtigkeit, Verlässlichkeit und Nachhaltigkeit anhand von Kriterienkatalogen konkretisieren und dadurch messbar und bewertbar machen, liegen vor (Hallensleben et al. 2020). Die gesellschaftliche Auseinandersetzung mit derartigen Konzepten hat jedoch erst begonnen. Etabliert sind solche Siegel bisher nicht.

Für algorithmische Systeme mit hohem Risiko sei laut EK (2020d) ein neuer Rechtsrahmen erforderlich: Bei der Risikoeinstufung sollen explizit zu benennende Anwendungsbereiche und das spezifische Risiko bei der Anwen-

dung einzelner Systeme berücksichtigt werden. Für diese KI-Systeme sollen unterschiedliche Maßnahmen zum einen deren Transparenz und Nachvollziehbarkeit sichern (u. a. durch die Verwendung von Trainingsdaten die europäischen Normen entsprechen und EU-repräsentativ sind; durch die Aufbewahrung von Daten und Aufzeichnungen des Trainingsprozesses, durch Informationspflichten zum Einsatz). Zum anderen sollen deren Genauigkeit und Robustheit bei der Entwicklung und Anwendung überwacht und gesichert werden (u. a. sollen Fehler angemessen bewältigt werden können und der Genauigkeitsgrad während der gesamten Einsatzphase überwacht und). Auch eine Form der menschlichen Aufsicht solle vorsehen werden. Dazu soll eine vorab durchzuführende Konformitätsbewertung durchgeführt werden, die Prüfung, Zertifizierung und Inspektionen umfasst sowie kontinuierliche Marktüberwachungen durch Behörden vorsieht. Haftungsfragen und Rechtsbehelfe im Schadensfall sollten separat definiert werden. Dieser Ansatz knüpft auf der Produktebene an Verfahren des Medizinprodukterechts mit den unterschiedlichen Risikoklassen und den Sicherheits- und Leistungsnachweisen an (ausführlicher in Kap. 4.2 u. 4.3).

Auch die nationale Datenethikkommission empfiehlt bei der Weiterentwicklung des Rechtsrahmens einen risikoadaptierten Regulierungsansatz, der auf Schadensvermeidung sowohl auf individueller als auch auf überindividueller Ebene zielt (DEK 2019, S. 173 ff.). Sie spricht sich dafür aus, neben den rechtlich bereits verankerten finanziellen und physischen Schäden auf individueller Ebene auch immaterielle Schäden unterschiedlicher Art, vor allem Grundrechtsverletzungen (von Teilhabebeschränkungen bis Diskriminierung) zu berücksichtigen. Das Schädigungspotenzial bzw. die Kritikalität solcher algorithmischen Systeme solle anhand der Schwere und der Eintrittswahrscheinlichkeit eines Schadens bestimmt werden. Unterschiedliche Kriterien wie Schadenshöhe, Betroffenenzahl, Reversibilität und Folgedimensionen von Entscheidungen sowie die Marktmacht der jeweils einsetzenden Stelle sollen zur Bewertung der Kritikalität solcher Systeme herangezogen werden. Je größer das Schädigungspotenzial solcher Systeme, desto stärkere Regulierungsmaßnahmen sollten in Betracht gezogen werden. Die Datenethikkommission schlägt eine fünfstufige Regulierung vor (Tab. 3.1). Die in der 5. Stufe befindlichen unvertretbaren Datenanalysen u. a. zu Profilbildungen, Totalüberwachung, Vulnerabilitäten gezielt ausnutzende Geschäftsmodelle oder Formen des Datenhandels sollten konkret benannt und explizit verboten werden.

Tab. 3.1 Kritikalitätsstufen algorithmischer Systeme

Stufe	Schädigungspotenzial	Regulierung
1	gering	keine speziellen Maßnahmen
2	leicht erhöht	formelle Anforderungen (Transparenzpflichten, Veröffentlichung einer Risikofolgenabschätzung) Ex-post Kontrollen
3	regelmäßig, deutlich	zusätzlich Ex-ante-Zulassung
4	erheblich	zusätzlich Liveschnittstelle zur kontinuierlichen Kontrolle durch Aufsichtsinstitution
5	unvertretbar	Verbot

Quelle: DEK 2019, S. 177

Neben der Arbeit an einer spezifischen *Verordnung zur Festlegung harmonisierter Vorschriften für KI* (EK 2021), die auf eine hohe (Produkt-)Sicherheit in Bezug auf lernende algorithmische Systeme abzielt, werden auf EU-Ebene auch die Verordnungen über digitale Märkte und digitale Dienste als übergeordnete Maßnahmen im Rahmen der Digitalstrategie der Europäischen Kommission verhandelt (Kasten 3.2). Sie sollen perspektivisch vielfältige digital relevante Regelungen aus unterschiedlichen Rechtsbereichen zusammenführen, harmonisieren sowie bestehende Regelungen ergänzen und Schutzlücken möglichst schließen. Zu den Verhandlungspunkten gehört auch die Weiterentwicklung der Regularien zum Umgang mit algorithmischen Systemen. Tendenziell verbrauchernahe und grundrechteschützende Akteur/innen weisen darauf hin, dass mit den etablierten Regularien vor allem bei Onlineaktivitäten weder Profilbildungen anhand von individuellen digitalen Datenspuren und umfangreichen Verkehrsdaten noch darauf aufbauende personalisierte Preisgestaltungen, Werbung und (Des-)Information adressiert und begrenzt werden können. Sie fordern strengere Verfahren zur Kontrolle algorithmischer Systeme bis hin zu expliziten Verboten. Auch eine europäische Agentur mit speziellen Kontroll- und Anordnungsbefugnissen wird im Rahmen der Verhandlungen zur Verordnung über digitale Dienste diskutiert. Große plattformbetreibende Akteure betonen vor allem Geschäftsgeheimnisse, Datenschutz und Risiken durch Algorithmenbeeinflussung, um Befugnisse von Aufsichts- und Kontrollgremien zu begrenzen. Die Abstimmungsprozesse sind noch nicht abgeschlossen (Stand November 2021).[74]

74 Zum Stand der Verfahren https://ec.europa.eu/digital-single-market/en/digital-services-act-package; https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-markets-act-ensuring-fair-and-open-digital-markets_de (10.11.2021)

Parallel zu den Überlegungen, den Rechtsrahmen zur Entwicklung und zum Einsatz von algorithmischen Systemen neu zu gestalten, gibt es auch Ansätze, die unterschiedliche Data-Mining-relevante Personengruppen in den Blick nehmen und deren spezifische Kompetenzen fördern wollen. Dazu gehört die stärkere Sensibilisierung derjenigen, die algorithmische Systeme entwickeln, sich prospektiv stärker mit möglichen Folgedimensionen auseinanderzusetzen. In diesem Kontext werden bereichsbezogene ethische Leitlinien und Verhaltensregeln genannt. Mitunter wird auf die in der Medizin seit jeher verankerten medizinethischen Grundprinzipien zur Schadensvermeidung, zu Patientenwohl und -autonomie sowie zu sozialer Gerechtigkeit verwiesen (Beauchamp/Childress 2008). Diese sind seit langem in der Ausbildung, im Berufsrecht aber auch bei der Prüfung von Anträgen zu klinischen Studien sowie sekundären Datenanalysen verankert und wirken sich auch auf Verfahren zum Datenzugang (regelmäßig über Treuhandstrukturen) sowie die Entwicklung datenanalytischer Verfahren aus (Sicherheits- und Leistungsnachweise für Medizinprodukte). Auch die Kompetenzen derjenigen seien zu stärken, die algorithmische Systeme zur Entscheidungsfindung einsetzen (teilweise mit englischen Begriffen wie Digital/Algorithm Literacy umschrieben). Zumindest ein Grundverständnis sowohl zur Funktionsweise der jeweiligen algorithmischen Systeme als auch zur Einordnung der ermittelten Ergebnisse sei regelmäßig nötig, um die Richtigkeit und Aussagekraft der Ergebnisse bewerten sowie mögliche Fehler erkennen zu können (Scholz et al. 2021, S. 110). Diese Kenntnisse werden bei steigendem Schädigungspotenzial der Resultate algorithmischer Systeme wichtiger. Auch diesbezüglich liefert der medizinische Bereich Anknüpfungspunkte. Dort ist sowohl der Einsatz vielfältiger Medizinprodukte als auch die Interpretation von Daten und Analyseergebnissen und die Ableitung von Therapieempfehlungen zertifiziertem Personal vorbehalten.

Die nachfolgende Fallstudie zum Einsatz von Data-Mining in der Medizin stellt die bestehenden Möglichkeiten und Grenzen der Entwicklung und des Einsatzes algorithmischer Systeme zur Entscheidungsunterstützung anhand unterschiedlicher Anwendungsbeispiele vor.

4 Data-Mining in der Medizin

Data-Mining in der Medizin steht für die Analyse von gesundheits- bzw. krankheitsbezogenen Daten sowie die datenbasierte Entwicklung von Entscheidungsregeln und algorithmischen Systemen, die medizinische Ziele haben: Krankheiten erkennen und diagnostizieren, den weiteren Krankheitsverlauf prognostizieren sowie die Behandlung optimieren und realisieren. Derartige Tätigkeiten fallen bisher in den Verantwortungsbereich von Ärzt/innen, die zur umfangreichen Dokumentation und Datenspeicherung verpflichtet sind. Nachfolgend werden zunächst rechtliche und technische Aspekte der Erfassung und der primären Speicherung medizinischer Daten angesprochen (Kap. 4.1). Wenn anhand derartiger Daten Strukturen und verallgemeinerbare Regeln gesucht sowie mathematisch-statistische Modelle angepasst werden und solche Data-Mining-Prozesse zu softwarebasierten algorithmischen Systemen führen, die Diagnose- und Behandlungsentscheidungen unterstützen können, unterliegt diese Software dem Medizinprodukterecht, das eine hohe Produktsicherheit gewährleisten soll. Da sich trotz hoher Sicherheitsstandards beim Einsatz algorithmischer Systeme fehlerhafte Ergebnisse und in Folge Schäden nie gänzlich ausschließen lassen, stellen sich auch Haftungsfragen (Kap. 4.2). Beispielhaft werden der Stand der Entwicklung datenbasierter algorithmischer Systeme für medizinische Aufgaben mit unterschiedlicher Tragweite, deren Wege in die Anwendung zur Diagnose und Behandlung von Krankheiten sowie damit ggf. verbundene spezifische Herausforderungen dargestellt (Kap 4.3).

Ärzt/innen bzw. medizinische Einrichtungen sind zudem an vielfältigen weiteren gesundheitssystemischen Prozessen beteiligt (Aufgaben im öffentlichen Interesse). Dafür müssen sie aus ihren Behandlungsdokumentationen definierte Datensätze ableiten und an bestimmte gesundheitssystemische Institutionen übermitteln (Kap. 4.4). Es entstehen unterschiedliche gesundheitssystemische Datenflüsse und bei einzelnen Institutionen kontinuierlich größer werdende Versorgungsdatenbestände, denen ebenfalls erhebliches Data-Mining-Potenzial unterstellt wird (dazu dann Kapitel 5).

4.1 Medizinische Daten: rechtliche und technische Aspekte

Bei der Erhebung medizinischer (Gesundheits-)Daten und jeglicher Verarbeitung haben Ärzt/innen (Kap. 4.1.1) und medizinische Einrichtungen (Kap. 4.1.2) Schlüsselpositionen. Sie sind die primären datenverarbeitenden Personen bzw. Stellen. Von der Datenverarbeitung betroffene Personen sind Patient/innen. Datenverarbeitende und betroffene Personen haben ein besonderes Schutz- und Abhängigkeitsverhältnis zueinander (Arzt-Patienten-Verhältnis). Ein Grund-

verständnis dieser Positionen und der gesetzlich definierten Rechte und Pflichten ist nötig, um Datenstrukturen bei medizinischen Behandlungsprozessen sowie Möglichkeiten und Grenzen der Weiterverwendung erschließen zu können. Unterschiedliche Initiativen zielen darauf ab, medizinische Primärdaten für komplexe Analysen zugänglicher zu machen (Kap. 4.1.3 u. 4.1.4).

4.1.1 Ärztinnen und Ärzte: Aufgaben, Pflichten, Aktenführung

Ärzt/innen haben eine Schlüsselposition bei der Behandlung von Patient/innen, bei der Erhebung und Verarbeitung medizinischer Daten aber auch bei vielfältigen gesundheitssystemischen Aufgaben. Der Arztberuf ist in Deutschland ein freier Beruf.[75] Um Patient/innen behandeln zu dürfen, müssen Ärzt/innen eine spezifische fachliche Qualifikation nachweisen und in einer Ärztekammer organisiert sein (fachliche Zulassung). Durch diese Kammerzugehörigkeit erhalten sie

- eine Zulassung zur Berufsausübung (Approbation) und unterliegen einer spezifischen Berufsordnung, die berufsrechtliche Pflichten und Leitlinien definiert;[76]
- eine deutschlandweit gültige, eindeutige Identifikationsnummer (die bundeseinheitliche Arztnummer – BAN), egal ob sie ambulant oder stationär praktizieren sowie
- einen elektronischen Heilberufsausweis (eHBA), mit dem sie sich authentifizieren und Datenverarbeitungsprozesse (rechts)sicher durchführen (neben der analogen und elektronischen Ausweisfunktion werden qualifizierte elektronische Signaturen, Datenverschlüsselungen und Zugriffsberechtigungen auf Behandlungsdaten damit realisiert).[77]

Zudem führen die Ärztekammern ein vollständiges Register mit den Namen, Zulassungsdaten und Nummern aller in Deutschland zugelassenen Ärzt/innen. Ambulant tätige Ärzt/innen, die medizinische Leistungen im Rahmen der gesetzlichen Krankenversicherung (GKV) abrechnen wollen, müssen sich zusätzlich in Kassenärztlichen Vereinigungen (KVen) organisieren (Kap. 5.3). Diese Vereinigungen auf Landesebene vergeben eine zweite Identifikationsnummer

75 Freiberufler erbringen eigenverantwortlich und fachlich unabhängig Dienstleistungen höherer Art im Interesse von Auftraggebern und der Allgemeinheit.

76 Das Standesrecht medizinischer Berufe ist föderal organisiert. Die meisten Landesärztekammern haben jedoch die von der Bundesärztekammer herausgegebene (Muster-)Berufsordnung der in Deutschland tätigen Ärztinnen und Ärzte (MBO-Ä) übernommen (Schepers et al. 2015, S. 230 ff.).

77 Etliche andere Heilberufler/innen sind ebenfalls in fachspezifischen Berufskammern organisiert und registriert (z. B. Hebammen, Physiotherapeut/innen), jedoch nicht alle (z. B. Notfallsanitäter/innen). Sie erhalten bisher keine eHBA (Bundesregierung 2018b).

(die lebenslange Arztnummer – LANR) und führen ein spezifisches Mitgliederregister mit Stammdaten und arztspezifischen Zusatzinformationen (KBV 2021). Stationär tätige Ärzt/innen sollen ebenfalls schrittweise eine weitere Arztnummer erhalten und die Deutsche Krankenhausgesellschaft ein bundesweites Register aller in Krankenhäusern und deren Ambulanzen tätigen Ärzt/innen einrichten (§ 108 SGB V). Diese Nummern, Ausweise und Register, die alle in Deutschland tätigen Ärzt/innen eindeutig identifizieren, sind bei vielen digitalen, datenbezogenen Prozessen im nationalen Gesundheitssystem zunehmend relevant (u. a. individuelle Signierung zahlreicher datenbezogener Prozesse, Verschlüsselung, Arztpseudonymisierung) und können insbesondere arztbezogene Datenanalysen ermöglichen.

Ärztliche Tätigkeiten, Pflichten und Haftung

Durch die Anerkennung der Berufsordnung verpflichten sich Ärzt/innen der Gesundheit des einzelnen Menschen und der Gesundheit der Bevölkerung zu dienen. Daraus ergibt sich zum einen ein Behandlungsauftrag (Anwendung des existierenden medizinischen Wissens). Zum anderen kann auch ein Forschungsauftrag (Wissenserweiterung) abgeleitet werden, der die Forschungsfreiheit als bürgerliches Grundrecht (Art. 5 Abs. 3 GG) für Ärzt/innen standesrechtlich anspricht. Sie verpflichten sich zudem zur gewissenhaften Berufsausübung (Sorgfaltspflicht). Dazu gehört u. a., dass sie behandlungsnotwendige medizinische Daten erheben und analysieren, sowie Daten Dritter im Zweifelsfall prüfen (Problematik: Fremdbefundung [Bundesregierung 2018d]), dass sie patientenbezogene Daten schützen und Schweigepflichten (Kasten 4.1) einhalten. Zur gewissenhaften Berufsausübung gehört auch, dass sie den anerkannten Stand des medizinischen Wissens beachten, geeignete Untersuchungs- und Behandlungsmethoden anwenden und Ergebnisse prüfen. Dazu kann auch der Einsatz algorithmischer Assistenzsysteme gehören.

Kasten 4.1 Ärztliche Schweigepflicht und medizinische Daten

Die Schweigepflicht ist sowohl straf- als auch berufsrechtlich definiert. Strafrechtlich sind Ärzt/innen und berufsmäßig tätige Gehilfen verpflichtet, über alle Dinge, die ihnen Patient/innen anvertraut haben (sowohl aus persönlichen, als auch aus betrieblichen oder geschäftlichen Lebensbereichen), zu schweigen, auch über deren Tod hinaus (§ 203 StGB). Eine unbefugte Offenbarung (auch gegenüber anderen Ärzt/innen) kann strafrechtlich mit Geld- oder Freiheitsstrafe sowie berufsrechtlich mit Geldbuße bis zum Entzug der Approbation (je nach landesrechtlichen Kammergesetzen) sanktioniert werden. Ein tatsächlicher Schaden bei Betroffenen muss damit nicht

verbunden sein. Die Befugnis zur Offenbarung können nur Patient/innen erteilen. Sie muss auf freier Entscheidung beruhen, d. h., Betroffene müssen abschätzen können, welche Daten warum an wen übermittelt werden. Darüber hinaus gilt eine Offenbarung als gerechtfertigt,

- wenn ein mutmaßliches Einverständnis angenommen werden kann (z. B. in Notfällen oder wenn mehrere Ärzt/innen eine Person gemeinsam behandeln, dann wird standardmäßig angenommen, dass sie untereinander von der Schweigepflicht entbunden sind und einer gemeinsamen Schweigepflicht unterliegen) und
- wenn zum Schutz eines höherwertigen Rechtsguts Ärzt/innen zur Auskunft gesetzlich verpflichtet werden (z. B. laut Infektionsschutzgesetz[78] beim Auftreten ansteckender Krankheiten oder laut SGB V zur Leistungsabrechnung; Kap. 4.4).

Rechtlich schulden Ärzt/innen ihren Patient/innen eine dem anerkannten Stand der medizinischen Wissenschaft entsprechende Behandlung (Rechtsgrundlage mündlicher Behandlungsvertrag [§ 630a ff. BGB]). Zu dieser gehören eine Informationspflicht über das Vorgehen und eine Dokumentationspflicht zum Vorgehen, jedoch kein Behandlungserfolg. Bei der Behandlung haben sie eine gewisse Therapiefreiheit. Sie können auch neue Medizinprodukte bzw. neue datenanalytische Verfahren einsetzen, sofern diese zertifiziert, sachlich gerechtfertigt und medizinisch vertretbar sind, Patient/innen darüber aufgeklärt werden, diese einwilligen und Ärzt/innen den Einsatz dokumentieren. Medizinische Fachgesellschaften geben zunehmend Leitlinien zur Behandlung heraus, in denen sie den medizinischen Wissenstand zu krankheitsspezifischen Behandlungsverfahren zusammenstellen. Sie sind eine wichtige Orientierung, um die Vertretbarkeit eines Verfahrens abzuschätzen. Der Einsatz fachlich nicht vertretbarer Verfahren sowie unzureichende Dokumentationen, die Zweifel an der informierten Einwilligung zulassen, sind Sorgfaltspflichtverletzungen bzw. Behandlungsfehler. Treten infolgedessen gesundheitliche oder materielle Folgeschäden auf, können Ärzt/innen haftbar gemacht werden. Dafür muss eine geschädigte Person mit hoher Sicherheit nachweisen, dass die Sorgfaltspflichtverletzung den Schaden verursacht hat (Verschuldenshaftung).

Nicht zertifizierte, neuartige Verfahren können Ärzt/innen im Rahmen von genehmigungspflichtigen *klinischen Studien* einsetzen. Diese basieren auf definierten Prüfplänen, in denen strukturierte Datenerhebungen und -analysen sowie detaillierte Dokumentationen und Sicherheitsmaßnahmen definiert werden. Patient/innen müssen explizit in die Teilnahme und in die Verwendung ihrer

78 Gesetz zur Verhütung und Bekämpfung von Infektionskrankheiten beim Menschen (Infektionsschutzgesetz – IfSG)

Daten zu definierten Zwecken freiwillig und informiert einwilligen. Bei klinischen Studien, in denen neue Produkte und Verfahren getestet und entwickelt werden, haften Ärzt/innen und Hersteller gemeinsam bei auftretenden Schäden ursachenunabhängig (Gefährdungshaftung). Ärztliche Berufshaftpflichtversicherungen decken ein wachsendes Spektrum möglicher Sorgfaltspflichtverletzungen ab (WD 2016, S. 4).

Arztgeführte Patientenakten

Seit 2013 sind Ärzt/innen verpflichtet, Patient/innen sowohl über wesentliche Umstände ihrer Behandlung in verständlicher Weise zu informieren als auch das Vorgehen zu dokumentieren und eine vollständige Patientenakte zu führen (§ 630f BGB). In dieser *arztgeführten Patientenakte* sind sämtliche aus fachlicher Sicht für die derzeitige und künftige Behandlung wesentlichen Maßnahmen und deren Ergebnisse aufzuzeichnen, insbesondere Anamnese, Untersuchungen, Ergebnisse, Befunde, Diagnosen, Eingriffe, Therapien und deren Wirkungen sowie diesbezüglich notwendige Aufklärungen und Einwilligungen. Auch Arztbriefe, Überweisungsschreiben und andere Dokumente von und an andere Ärzt/innen sind aufzunehmen. Nachträgliche Änderungen müssen kenntlich gemacht werden und der ursprüngliche Inhalt erkennbar bleiben. Der primäre Zweck dieser Akte ist die Dokumentation einer sachgerechten Behandlung (Bundesregierung 2012b, S. 25 f.). Wichtige Daten und Informationen sollen aufbewahrt werden, um die Rechte von Patient/innen (insbesondere deren Recht auf Rechenschaft gegenüber behandelnden Ärzt/innen) zu wahren, die Behandlungssituationen nachvollziehen und gegebenenfalls Beweislastfragen in Haftungsfällen klären zu können. Bei dieser Klärung gilt inzwischen: Was in der arztgeführten primären Patientenakte nicht nachvollziehbar dokumentiert ist, hat nicht stattgefunden (Krokotsch 2017). Die Dokumentation erfolgt ausschließlich durch behandelnde Ärzt/innen. Sie müssen dafür weder Einwilligungen bei ihren Patient/innen einholen, noch diese explizit darüber informieren. Patient/innen haben jedoch Einsichts- und Abschriftsrechte. Die ärztliche Dokumentationspflicht beschränkt einerseits das informationelle Selbstbestimmungsrecht von Patient/innen (Kap. 3.3.1). Andererseits unterliegen alle in der arztgeführten Akte enthaltenen Daten/Ergebnisse/Informationen der ärztlichen Schweigepflicht (Kasten 4.1).

Bei der Einführung der Dokumentationspflicht 2013 lagen die Klärung von Haftungsfragen und die Stärkung diesbezüglicher Patientenrechte im Fokus der Regulierung, nicht die mögliche Weiterverwendung der Akteninhalte für komplexe Datenanalysen bzw. Data-Mining. Es wurde keine mögliche sekundäre Nutzung der Akteninhalte im Rahmen der ärztlichen Schweigepflicht angesprochen, keine standardisierte, elektronische Datenspeicherung definiert. Sogar die

Speicherung der bei der Behandlung entstehenden (Roh-)Daten blieb im Ungefähren, lediglich Untersuchungsergebnisse und Befunde werden explizit als zu speichern benannt. Die Mindestaufbewahrungsfrist der Patientenakten beträgt 10 Jahre nach Behandlungsabschluss, teilweise ist sie länger (§ 630f BGB).[79] Löschfristen wurden nicht definiert. Diese rechtliche Lösung zur Klärung von Haftungsfragen (Dokumentation ohne Patienteneinwilligung unter Achtung von Schweigepflichten) führt dazu, dass in jeder medizinischen Einrichtung eigenständige Patientenakten geführt werden.

4.1.2 Medizinische Einrichtungen: Organisation und Datenverwaltung

Medizinische Einrichtungen sind die Geschäftseinheiten, in denen Patient/innen behandelt bzw. in denen diagnose- und behandlungsrelevante Spezialleistungen von Ärzt/innen und anderen medizinischen erbracht werden. Sie werden in Deutschland einem von zwei organisatorisch, verwaltungstechnisch und regulativ getrennten Sektoren zugeordnet:

- Der *stationäre Sektor* umfasst in Deutschland knapp 2.000 Krankenhäuser in unterschiedlicher Trägerschaft (öffentliche, konfessionelle, privatwirtschaftliche). Die Spanne reicht von Allgemeinkrankenhäusern mit Grund- und Regelversorgung über spezialisierte Fachkliniken bis zu (Universitäts-)Kliniken mit Schwerpunkt- und Maximalversorgung. Krankenhäuser beschäftigen vielfältige Fachkräfte und verfügen über eigene Arzneimittelversorgungsstrukturen. Krankenhausträger agieren rechtlich und wirtschaftlich selbstständig.
- Zum *ambulanten Sektor* gehören mehr als 100.000 Einzel- und Gemeinschaftspraxen niedergelassener Ärzt/innen sowie diverse weitere Einrichtungen, die bei der Diagnostik und bei Behandlungen mitwirken (u. a. Labore, Apotheken). Ambulante Einrichtungen werden wirtschaftlich betrieben. Patient/innen werden oft über lange Zeit fachspezifisch betreut.

Medizinische Einrichtungen werden im Rahmen der gesundheitssystemischen Selbstverwaltung zugelassen, registriert und überwacht. Die fachliche Zulassung von Krankenhäusern erfolgt über Krankenhausgesellschaften auf Landes- und Bundesebene. Für die Registrierung vergeben sie Kennzeichen für jede stationäre Einrichtung und führen ein Register mit vielfältigen Stammdaten und ergänzenden Informationen.[80] Im ambulanten Bereich vergeben Kassenärztli-

79 Transfusionsgesetz, Röntgen- und die Strahlenschutzverordnung sowie landesweite Krankenhausgesetze verlangen Akten bis zu 30 Jahre aufzubewahren (Schepers et al. 2015, S. 96).

80 www.deutsches-krankenhaus-verzeichnis.de/das-dkv/ueber-das-dkv (10.11.2021)

che Vereinigungen eindeutige Betriebsstättennummern (BSNR) und führen entsprechende Register mit definierten Stammdaten aller Arztpraxen, die sich an der medizinischen Versorgung im Rahmen der GKV beteiligen. Bei Apotheken übernehmen die Landesapothekenkammern die Kennzeichenvergabe und Registerführung. Jedes Register enthält definierte Stammdaten und Zusatzinformationen. Alle Einrichtungen, die Leistungen mit Sozialleistungsträgern direkt abrechnen, benötigen zusätzlich ein zweites eindeutig identifizierendes Institutionenkennzeichen (IK), das die Arbeitsgemeinschaft Institutionenkennzeichen vergibt. Sie führt ein eigenes Register, das ausschließlich abrechnungsrelevante Daten enthält (Name, Anschrift, Telefonnummern, Bankverbindung), die nur für Aktivitäten der sozialen Sicherung verwendet und Dritten nicht zugänglich gemacht werden (ARGE IK 2015, S. 33). Diese Kennzeichen werden auch als Einrichtungspseudonyme für standardisierte Leistungsabrechnungsdatensätze verwendet (Kap. 4.4.1, Abb. 4.4). Über diese Nummern, Kennzeichen und Register sind alle in Deutschland zugelassenen medizinischen Einrichtungen eindeutig identifizierbar. Sie ermöglichen vielfältige einrichtungsbezogene Datenanalysen. Die unterschiedlichen Register bilden zudem ein Datenfundament für die Entwicklung spezifischer Informationsdienste, mit denen registerführende Stellen ihre Serviceangebote erweitern können, sofern dies im Rahmen ihres Aufgabenbereichs liegt.

Praxis- und Krankenhausinformationssysteme

In medizinischen Einrichtungen werden datenverarbeitende Prozesse zunehmend mithilfe von Informationssystemen realisiert. Es gibt unterschiedliche fachspezifische (z. B. Radiologie- oder Laborinformationssysteme) sowie allgemeine Praxis- oder Krankenhausinformationssysteme (PIS/KIS).[81] In der Summe handelt es sich um einrichtungsspezifisch angepasste, weitgehend geschlossene Softwareinsellösungen, die bisher ausschließlich als proprietäre Software von unterschiedlichen Herstellern angeboten wird.[82] Diese Konstellation erschwert u. a. die Vernetzung von Systemen unterschiedlicher Anbieter, vollständige Systemwechsel sowie Softwareentwicklungen Dritter.

PIS/KIS haben unterschiedliche Arbeitsbereiche. Klinische Bereiche mit diversen fachspezifischen Modulen sind behandelnden Ärzt/innen vorbehalten und als Online-Transaction-Processing-System (OLTP) für Dateneingaben und

81 Es gibt vielfältige medizinische Informationssysteme, mit unterschiedlichen Bezeichnungen und inhaltlichen Schwerpunkten. Im Bericht wird vereinfachend von PIS/KIS gesprochen. Damit sind alle in medizinischen Einrichtungen eingesetzten datenverarbeitenden IT-Systeme gemeint.

82 Im stationären Bereich gibt es etwa 30 Systeme, im ambulanten Bereich mehr als 150. Einzelne Einrichtungen setzen Softwaresonderlösungen ein (ausführlicher z. B. Weichert 2018, S. 31 ff.)

-präsentationen optimiert (Schepers et al. 2015, S. 183). In diesen Arbeitsbereichen werden u. a. die primären Patientenakten geführt. In administrativen Arbeitsbereichen können verwaltungstechnische Aufgaben zunehmend digital realisiert werden. Dafür erhalten andere Fachkräfte z. T. selektive Leserechte, um definierte Datensätze für diverse administrative Aufgaben zusammenstellen zu können; Kap. 4.4).

PIS/KIS sind nicht dafür konzipiert, Daten aktenübergreifend zu analysieren, um Strukturen zu erkennen, Trainingsdaten abzuleiten und aufzubereiten oder Modelle zu trainieren. Anwenderzielgruppe sind Ärzt/innen und medizinische Fachkräfte, keine Datenanalysten (Schepers et al. 2015, S. 183). Sollen medizinische Behandlungsdaten für Data-Mining-Aktivitäten verwendet werden, müssten diese aus PIS/KIS exportiert werden. Entstehen bei diesen externen analytischen Prozessen algorithmische Assistenzsysteme, könnten diese im Anschluss als zusätzliche Softwaremodule in PIS/KIS integriert werden, um Ärzt/innen oder administrative Fachkräfte bei spezifischen Tätigkeiten zu unterstützen (z. B. bei der Klassifikation neuer Situationen oder Fälle).

Je mehr medizinische und administrative Prozesse mithilfe dieser Informationssysteme realisiert werden, desto gravierender sind Störungen. Die IT-Systeme großer Krankenhäuser gelten bereits seit einigen Jahren als in besonderem Maße zu schützende kritische Infrastruktur (Kasten 3.1). Ab 2022 sind alle stationären Einrichtungen, die Leistungen im Rahmen der gesetzlichen Krankenversicherung erbringen, verpflichtet, ihre IT-Systeme vor Cyberattacken in besonderem Maße zu sichern (§ 75c SGB V).

4.1.3 Medizinische Primärdaten

Eine allgemeingültige Strukturierung oder eindeutige Bezeichnung medizinisch relevanter Daten gibt es bisher nicht (Schepers et al. 2015, S. 85). Charakteristisch sind die personen-/patientenbezogene Erfassung sowie die gesundheitsbezogenen, medizinisch-fachlichen Inhalte. Nachfolgend werden derartige Daten anhand der Bedingungen, unter denen sie generiert und verwendet werden dürfen, unterschieden. Einerseits wird von Daten aus klinischen Studien und andererseits von Behandlungsdaten gesprochen. Mit der Behandlung sind im nationalen Gesundheitssystem vielfältige öffentliche Aufgaben verbunden, an denen Ärzt/innen und medizinische Einrichtungen beteiligt sind, und die sie als medizinische Leistungserbringer zur Ableitung und Übermittlung unterschiedlicher Daten gesetzlich verpflichten *(administrative Daten*; Kap. 4.4).

Daten klinischer Studien

Eine seit jeher wichtige Datenquelle für Data-Mining-Prozesse in der Medizin sind geplante klinische Studien. Spezifische Prüfpläne definieren die genaue

Vorgehensweise (Studienpopulation, zu erfassende Parameter, einzusetzende Messverfahren, anzuwendende Codierungen und Klassifikationen usw.). Die Standards der guten klinischen Praxis (DIN EN ISO 14155) erfordern eine Begutachtung der Prüfpläne durch eine Ethikkommission und ein Qualitätsmanagement bei der Studiendurchführung. Die Ethikkommission wägt die mit einer Studie verbundenen gesundheitlichen Risiken für Teilnehmende mit dem angestrebten Gesamtnutzen der Studie ab. Jegliche Teilnahme an klinischen Studien ist freiwillig. Teilnehmende willigen in die Datennutzung zu den in der Regel weit definierten Forschungszwecken ein (auch als Broad Consent bezeichnet). Rechtsgrundlage sind weitgehend standardisierte Nutzungslizenzen. Die Vorgaben der DSGVO gelten vollumfänglich (Kap. 3.3 ff.).

Es gibt unterschiedliche Studienkonzepte (ausführlich z. B. TAB 2010). Bei *Interventionsstudien* werden quasi Behandlungen nach Plan durchgeführt und unterschiedliche Konzepte miteinander verglichen. Mit diesen werden vor allem die Sicherheit und Wirksamkeit bzw. Leistungsfähigkeit von Arzneimitteln, Medizinprodukten oder Behandlungsverfahren geprüft (Kap. 4.2). Im Rahmen der Entwicklung von Arzneimitteln aber auch von Medizinprodukten hoher Risikoklassen sind sie verpflichtend und folglich oftmals industriefinanziert (auch als kommerzielle Studien bezeichnet). Den damit generierten Datenbestand kann der Studiensponsor bisher monopolisiert verwenden, was seit Jahren kontrovers diskutiert wird (Kasten 8.1). Bei *beobachtenden Studien* werden vielfältige Daten zu teilnehmenden Personen entsprechend der jeweiligen Studienpläne erfasst, ohne vorzuschreiben, wie eine medizinische Behandlung im Krankheitsfall erfolgen soll. Mit Querschnittsstudien wird der Zustand teilnehmender Patient/innen situativ zunehmend detailliert erfasst, wie z. B. bei den weltweit koordinierten Aktivitäten zum Aufbau von Tumormutationsregistern (Kap. 4.3.3). Eine andere Form sind langfristige Beobachtungen größerer Personengruppen, teils auch als epidemiologische, Bevölkerungs- oder Längsschnittstudien bezeichnet. Die 2014 gestartete NAKO-Gesundheitsstudie (ehemals Nationale Kohorte)[83] ist die bisher größte nationale Längsschnittstudie. Nach derzeitiger Planung sollen 200.000 Menschen über einen Zeitraum von 20 bis 30 Jahren regelmäßig befragt, medizinisch untersucht und vielfältige gesundheitsbezogene Daten erhoben werden, um in diesen retrospektiv nach Ursachen und Risikofaktoren vielfältiger (Volks-)Krankheiten zu suchen. Derartige Studien werden in Deutschland in erheblichem Maße öffentlich finanziert, im Rahmen der Forschungsförderung fachlich und ethisch geprüft und von Forschungseinrichtungen durchgeführt. Die entstehenden Datenbestände können zunehmend über Treuhandstrukturen entsprechend definierter Zugangsregeln genutzt werden (Kap. 3.3.3).

83 https://nako.de/allgemeines/was-ist-die-nako-gesundheitsstudie/ (10.11.2021)

Da die Datenanalyse explizites Ziel klinischer Studien ist, werden zum einen entsprechende breite Einwilligungen bei teilnehmenden Personen bereits bei der Studienteilnahme eingeholt und zum anderen Daten entsprechend der Prüfpläne strukturiert erhoben und in maschinenlesbaren Formaten in spezifischen Repositorien gespeichert (generierte Datensätze und -bestände haben im Vergleich zu normalen Behandlungsdokumentationen eine vergleichsweise gute semantische und syntaktische Interoperabilität). Daten aus klinischen Studien haben einige methodisch/inhaltliche Spezifika, die bei der Weiterverwendung u. a. für Data-Mining berücksichtigt werden sollten: In klinische Studien werden je nach Untersuchungsfrage nur bestimmte mehr oder weniger große Patientengruppen eingeschlossen. Einige Personengruppen sind oftmals unterrepräsentiert oder gänzlich ausgeschlossen (u. a. Kinder, ältere oder multimorbide Personen). Dies beschränkt mitunter die Repräsentativität von Studiendaten. Aus statistischer Perspektive handelt es sich bei klinischen Studien immer um mehr oder weniger große Stichproben (keine Totalerhebungen). Zudem gelten medizinische Daten aus Interventionsstudien als unter Laborbedingungen generiert. Mitunter gibt es Abweichungen zu Daten, die unter Alltagsbedingungen (z. B. im Rahmen der allgemeinen medizinischen Behandlung) generiert werden.

Medizinische Behandlungsdaten in arztgeführten Primärakten

Medizinische Daten, die im Rahmen der regulären Behandlung generiert werden, sind weniger standardisiert als Studiendaten. Zum einen wägen Ärzt/innen eigenverantwortlich das situativ notwendige medizinische Vorgehen in Abstimmung mit den Betroffenen ab (Therapiefreiheit unter Achtung des anerkannten Wissens, kein definiertes Vorgehen laut Prüfplan). Zum anderen gibt es bisher kaum normative Vorgaben zur Art und Weise der Aktenführung. Dennoch liegt es nahezu auf der Hand, dass die seit 2013 in Patientenakten gesammelten und ohnehin lange Zeit aufzubewahrenden medizinischen Behandlungsdaten Fragen bezüglich der Weiterverwendung und Analyse aufwerfen. Denn in der Summe bilden diese Daten das medizinische Behandlungsgeschehen in Deutschland hochgranular und weitgehend vollständig ab (Totalerhebung). Es gibt jedoch technische Barrieren und rechtliche Vorgaben, die die Weiterverwendung für Data-Mining-Aktivitäten begrenzen und lenken. Zudem müssen einige methodisch/inhaltliche Spezifika beachtet werden.

Die bisherige Rechtskonstruktion führt dazu, dass in jeder medizinischen Einrichtung eigenständige Patientenakten geführt werden. Da niedergelassene Ärzt/innen ihre Patient/innen oft über längere Zeiträume betreuen, enthalten deren Primärakten kontinuierliche Dokumentationen, die jedoch fachspezifisch begrenzt sind (bei der hausärztlichen Versorgung entstehen andere Behand-

lungsdaten als z. B. bei der zahnärztlichen). Für Spezialdiagnostiken und -behandlungen wird teilweise Probenmaterial verschickt oder es werden Patient/innen an andere Einrichtungen überwiesen, die z. B. spezielle labordiagnostische, genetische, radiologische (Roh-)Daten erheben und oftmals auch interpretieren und befunden. Im Rahmen beauftragter Untersuchungen oder gemeinschaftlicher Behandlung werden oft nur auffällige Untersuchungsergebnisse (Befunde) u. a. mittels Arztbriefen oder Entlass-/Überweisungsdokumenten verbal beschrieben, fachspezifisch erläutert und zurückgemeldet, nicht aber alle (Roh-)Daten, die u. a. durch genetische Untersuchungen, bildgebende Verfahren usw. erhoben wurden. In Krankenhäusern wird das Befinden von Patient/innen zwar detailliert, aber nur während eines kurzen Zeitraums erfasst. In der Summe werden das Befinden und die medizinische Behandlung von Patient/innen fragmentiert dokumentiert, in ambulanten Einrichtungen eher anhand fachspezifisch begrenzter Längsschnittdaten und in stationären Einrichtungen anhand vielfältiger Querschnittsdaten (Schepers et al. 2015, S. 130).

Die Inhalte der arztgeführten Primärakten sind rechtlich durch Datenschutz- und Schweigepflichten geschützt und technisch weder behandelnden Ärzt/innen anderer Einrichtungen noch anderen Dritten unmittelbar zugänglich. Die verteilte und gekapselte Haltung von Behandlungsdaten in unterschiedlichen Akten prägt sowohl die Behandlungsprozesse als auch die Kommunikation zwischen Ärzt/innen in unterschiedlichen Einrichtungen (Stichwort: Mehrfachuntersuchungen und -datenerfassungen). Das wesentliche Kommunikationsmedium zwischen medizinischen Einrichtungen sind derzeit textbasierte Dokumente (Arztbriefe, Überweisungs- und Entlassdokumente sowie Rezepte). Darin enthaltene Daten und Informationen sind an Fachkräfte adressiert, überwiegend fachsprachlich formuliert und als Freitext aufgezeichnet (Ausnahme standardisierte Rezept-, Überweisungs- und Arbeitsunfähigkeitsbescheinigungen). In die Patientenakten empfangender Einrichtungen werden diese Dokumente oft nur als Bild- oder Textdatei aufgenommen, nicht aber in maschinenlesbaren Formaten gespeichert. In der Summe haben die arztgeführten Patientenakten einen geringen Interoperabilitätsgrad. Erst ansatzweise werden standardisierte Terminologien und Formulare eingesetzt (Kasten 4.2).

Kasten 4.2 Standardisierung und Interoperabilität in der Medizin

Interoperabilität bezeichnet die technische Möglichkeit, Daten systemübergreifend verarbeiten zu können (Schepers et al. 2015, S. 126 ff.). Kernelement sind gemeinsame Standards. Auf vielfältige Art und Weise wird daran gearbeitet.

- Harmonisierung von Terminologien und Klassifikationen *(semantische Interoperabilität)*: Es gibt diverse fachspezifische medizinische Klassi-

fikationen (z. B. die radiologische Befundklassifikation für Mammografieaufnahmen; Abb. 3.1) und Aktivitäten, diese zu einer umfassenden medizinischen Terminologie zusammenzuführen. Diesbezüglich setzt sich die Systematized Nomenclature of Medicine (SNOMED) zunehmend durch, die mit mehreren Hunderttausend Begriffen und Konzepten medizinische Aussagen eindeutig und vollständig abbilden will.

- Standardisierung von Datenformaten, Dokumentationen und Formularen (*syntaktische Interoperabilität*): Es gibt bereits etliche informationstechnische Standards z. B. für medizinische Bilder den Digital Imaging and Communications in Medicine (DICOM), der objektorientierte Datenbankmodelle (Kap. II.2.2.1) nutzt, oder für textbasierte klinische Dokumente die Clinical Document Architecture (CDA), sie nutzt die erweiterbare textbasierte Auszeichnungssprache Extensible Markup Language (XML), die die maschinelle Verarbeitung erleichtert. Die Art und Weise der Dokumentation und die Entwicklung standardisierter Formulare sind jedoch eine große Herausforderung.

Zudem gibt es internationale Initiativen, die semantische und syntaktische Interoperabilität zwischen medizinischen Bereichen fördern, z. B. Integrating the Healthcare Enterprise (IHE). Andere wollen darüber hinaus auch den Datenaustausch zu anderen gesundheitssystemischen Einrichtungen (z. B. Krankenkassen, die in der Regel administrative/statistische Klassifikationen verwenden) verbessern, z. B. Health Level 7 (HL 7)[84]. Auf nationaler Ebene muss entschieden werden, welche Standards sowohl in einzelnen medizinischen Fachdisziplinen als auch gesundheitssystemisch verwendet werden sollen (Kasten 3.4), welche nationalen Besonderheiten und Anpassungen erforderlich sind und wie diese dann schrittweise mit den gesundheitssystemisch und historisch gewachsenen Strukturen zusammengefügt werden können. Auch dazu gibt es vielfältige Initiativen, z. B.

- arbeiten medizinische Fachgesellschaften an der Standardisierung fachspezifischer Prozesse, um einrichtungsübergreifend einheitlich zu kommunizieren (z. B. arbeitet die deutsche interdisziplinäre Vereinigung Intensiv- und Notfallmedizin [DIVI] an der Standardisierung minimaler Notfalldatensätze und an Rettungsdienstprotokollen, die in Notarztinformationssystemen zusammenlaufen sollen);
- stellt das Portal für Medizinische Datenmodelle elektronische Formulare für vielfältige medizinische Versorgungsprozesse in diversen Dateiformaten bereit, um Behandlungsdaten bereits strukturiert zu erfassen;[85]

84 Internationale Organisation zur Harmonisierung medizinischer und administrativer Klassifikationen und der Entwicklung interoperabler Standards für Gesundheitssysteme.
85 https://medical-data-models.org/ (10.11.2021)

- engagieren sich nationale IHE- und HL 7-Akteure in Bezug auf die Vereinheitlichung von elektronischen Medikationsplänen, Arztbriefen oder gar Patienten-/Fallakten, wie im Gesetz für sichere digitale Kommunikation und Anwendungen im Gesundheitswesen (E-Health-Gesetz) von 2015 gefordert und
- baut die gematik seit 2017 das elektronische Interoperabilitätsverzeichnis für technische und semantische Standards, Profile und Leitfäden für IT-Systeme im Gesundheitswesen aus.[86]

Sekundärnutzung zu Forschungszwecken: die rechtliche Situation

Sowohl die Datenschutz-Grundordnung als auch die ärztliche Schweigepflicht verbieten zwar die unerlaubte Offenbarung von patientenbezogenen Daten gegenüber anderen datenverarbeitenden Stellen, nicht aber die eigene/interne Verwendung zu Forschungszwecken. Vor allem bei großen medizinischen Einrichtungen entstehen durch die Dokumentations- und Aufbewahrungspflichten zunehmend große Datenbestände. Bei ihnen stellt sich seit Jahren vordringlich die Frage, wie weit der Rahmen der kollegialen Schweigepflicht gespannt werden kann, um medizinische Behandlungsdaten zu Forschungszwecken weiterzuverwenden. Unikliniken, die neben dem Versorgungs- auch einen expliziten Forschungsauftrag haben, thematisieren diese Sekundärnutzung medizinischer Behandlungsdaten in besonderem Maße. Die Rechtslage zur Weiterverwendung dieser Daten zu Forschungszwecken mit ihren Verflechtungen auf Bundes- und Landesebene bezüglich Schweigepflicht, Datenschutz und Datennutzung ist komplex: Unterschiedliche Rechtsnormen auf Landes- und Bundesebene legen den Forschungsbegriff unterschiedlich weit aus (teilweise wird nur für nichtkommerzielle Forschungsabsichten Zugang zu den Behandlungsdaten gewährt); verlangen unterschiedliche Datenschutzmaßnahmen (teilweise reicht Pseudonymisierung, teils wird Anonymisierung gefordert) und gewähren einem unterschiedlich großen Personenkreis den Zugang (teilweise nur Ärzt/innen einer Krankenhausabteilung, teilweise weitere Personen) (ausführlich z. B. Schneider 2015).

Aus der datenanalytischen Perspektive verlieren die Diskussionen zur Zulässigkeit der internen Datennutzung im Rahmen der Schweigepflicht ein Stück weit an Relevanz, denn besonderes Potenzial haben vor allem Datenbestände, die einrichtungsübergreifend nicht nur von behandelnden Ärzt/innen, sondern auch von Datenanalyst/innen weiterverwendet werden dürfen. Dazu sind Einwilligungen von Patient/innen in jedem Fall erforderlich. Dieser Weg wird zunehmend eingeschlagen (siehe Medizininformatik-Initiative).

86 www.vesta-gematik.de (10.11.2021)

Bei einrichtungsübergreifenden Datenverwendungen könnten urheber- bzw. leistungsschutzrechtliche Fragen an Bedeutung gewinnen. Derartige Fragen werden bislang kaum thematisiert. Auch diesbezüglich ist die Situation unübersichtlich. Sind z. B. Arztbriefe oder medizinische Bilder urheberrechtlich geschützt? Dürfen sie ohne Einwilligung der erstellenden Person jenseits von Dokumentationszwecken weiterwendet werden? Wie weit reicht die dort bisher temporär verankerte Wissenschaftsschranke (Kap. 8.1)?

Initiativen zur besseren Zugänglichkeit und Nutzbarkeit

Es gibt unterschiedliche Initiativen, die darauf abzielen, die Zugänglichkeit zu und Weiterverwendbarkeit von medizinischen Behandlungsdaten zu verbessern. Ohne Anspruch auf Vollständigkeit sollen einige für Data-Mining-Aktivitäten potenziell relevante Initiativen angesprochen werden. Seit Jahren wird national und international an Interoperabilitätsverbesserungen gearbeitet (Kasten 4.2). Aufgrund der kontinuierlichen Weiterentwicklung von Diagnose- und Behandlungsmöglichkeiten ist das eine Aufgabe, die dauerhaftes Engagement erfordert. Nach jahrelangen Diskussionen konnte inzwischen eine Einigung bezüglich der Harmonisierung von Terminologien erreicht werden. SNOMED soll zukünftig als nationale Referenznomenklatur für die Haltung von Behandlungsdaten in arztgeführten Primärakten verwendet werden (Bundesregierung 2020a, S. 7). Dabei wird der Blick stärker nach vorn gerichtet. Inwiefern bereits archivierte Behandlungsdaten aufbereitet und umformatiert werden, ist dagegen offen. Klar ist, dass eine solche Aufbereitung sehr aufwendig wäre. Nur wenn umfangreiche Weiterverwendungen archivierter Daten anvisiert werden, lohnt sich der Aufwand.

Seit etlichen Jahren ist die sichere Vernetzung der IT-Systeme unterschiedlicher medizinischer Einrichtungen mittels *Telematikinfrastruktur* (TI) ein gesundheitssystemischer Aktivitätsschwerpunkt. Die TI soll als interoperable Informations-, Kommunikations- und Sicherheitsinfrastruktur die IT-Systeme aller Beteiligten des Gesundheitswesens vernetzen – von medizinischen Einrichtungen über Kostenträger bis zu den Patient/innen bzw. Versicherten (§ 306 SGB V). Im Fokus steht die Verbesserung von Behandlungsprozessen. Dokumente sollen leichter datenschutzkonform elektronisch übertragen werden können. Ein direkter Zugang zu den arztgeführten Primärakten wird derzeit nicht diskutiert. Die Verbesserung komplexer datenanalytischer Aktivitäten wird bisher nicht explizit genannt, kann jedoch mittelbar durch die an die TI angeschlossenen patientengeführten Akten und die damit verbundenen elektronischen Kommunikationsmöglichkeiten mit den Patient/innen erwartet werden.

Debatten zum pro und kontra dezentraler Datenhaltung in Patientenakten werden in Deutschland seit Jahren geführt. 2003 gab es eine politische Initiative zur Einführung einrichtungsübergreifender elektronischer Krankenakten

(Schröder 2003). 2004 wurde im SGB V die Einführung einer *elektronischen Patientenakte* (ursprünglich als einrichtungsübergreifende arztgeführte Akte geplant) und einer *elektronischen Gesundheitskarte* (ursprünglich sowohl mit Ausweis- als auch mit Datenspeicherfunktionen für Versicherte geplant) verankert. Seitdem werden verschiedene Begriffe (Patienten-/Fall-/Kranken-/Gesundheitskarten/-akten/-konten oder -fächer) für gleiche oder ähnliche Konzepte sowie gleiche Begriffe für verschiedene Konzepte (bezüglich Datenhaltung [dezentral, zentral], Verantwortung [arzt- oder patientengeführt] sowie Inhalten und Zugänglichkeiten) verwendet und deren Möglichkeiten, Zwecke, Ziele aber auch Risiken (u. a. durch Cyberattacken) diskutiert. Bis heute gibt es unterschiedliche Ansichten, wie medizinische Behandlungsdaten bestmöglich geschützt, welche Funktionen Karten und Akten übernehmen und wie deren Inhalte unter Wahrung von Schweige-, Sorgfalts-, Dokumentations- und Haftpflichten für unterschiedliche Zwecke zugänglich gemacht werden können (ausführlich in Haas 2017).

Nach derzeitigem Planungsstand bleiben die dezentralen, arztgeführten Patienten-/Fallakten weiterhin die primären Akten für die medizinische Behandlung und deren Dokumentation. Zusätzlich sind gesetzliche Krankenkassen seit 2021 verpflichtet, ihren Versicherten eine elektronische Patientenakte als freiwilliges, sekundäres, versichertengeführtes, individuell verschlüsseltes Datenverwaltungssystem anzubieten (§ 341 ff. SGB V). Damit soll das Recht auf Datenübertragbarkeit für Patient/innen und die Pflicht zur Datenbereitstellung für Ärzt/innen schrittweise elektronisch realisiert werden können. Nach und nach sollen unterschiedliche medizinische Behandlungsdaten und Dokumente in standardisierter Form in diese sekundären Akten übertragen werden können. Dazu müssen jedoch die primären arztgeführten Patientenakten weiterentwickelt und deren Interoperabilität erhöht werden. Zudem soll es auch möglich werden, dass Versicherte eigene Dokumente und selbst erhobene Vitaldaten sowie die von ihrer Krankenkasse gespeicherten versichertenbezogenen Daten einstellen können. Mit der zur Aktenführung notwendigen Software wird erstmalig eine elektronische Kommunikationsmöglichkeit für und mit Patient/innen geschaffen. Auch ein persönliches Datenmanagementsystem soll schrittweise etabliert werden. Nach derzeitigem Planungsstand sollen Patient/innen ab 2022 selektiv in einrichtungs-, fach- und sektorübergreifende Datenübertragungen für medizinische Behandlungs- und gesundheitliche Versorgungszwecke einwilligen können. Sie sollen auch die Möglichkeit bekommen, ihren behandelnden Einrichtungen die Forschung mit ihren Daten breit zu erlauben. Perspektivisch könnten forschende medizinische Einrichtungen ihre derzeit noch papierbasierten Einwilligungsverfahren über diesen Kommunikationskanal organisieren. Nach derzeitigem Planungsstand sollen Versicherte spätestens ab 2023 die Daten ihrer Akte als formal anonymisierte Mikrodaten auch allgemein zu wissen-

schaftlichen Forschungszwecken freigeben können (auch als Datenspende bezeichnet). Diese Daten sollen dann an das Forschungsdatenzentrum der GKV übermittelt werden, das vom Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM) in fremdnütziger Treuhandschaft betrieben wird (Kap. 5.5.2).[87] Damit sollen Personen, die selbst eine Patientenakte führen möchten, ihre Daten auch dann zu Forschungszwecken spenden können, wenn sie in medizinischen Einrichtungen behandelt werden, die selbst keine datenbasierte Forschung betreiben.

Diese patientengeführte elektronische Patientenakte hat ein besonderes Data-Mining-Potenzial, weil zusätzlich zu den medizinischen Behandlungsdaten perspektivisch auch die bei den Krankenkassen gespeicherten Daten sowie individuell erhobene Vitaldaten personenbezogen zusammengeführt und für Forschungsaktivitäten zugänglich gemacht werden können. Ob dieses Konzept der versichertengeführten Haltung und Zusammenführung vielfältiger gesundheitsbezogener Daten von den Versicherten bzw. Patient/innen angenommen wird und sich die Potenziale tatsächlich realisieren lassen, kann gegenwärtig nicht seriös eingeschätzt werden. Denn viele Fragen sind zum jetzigen Zeitpunkt offen: Wie lange dauert es, bis medizinische Behandlungsdaten in interoperablen Formaten in die Patientenakten eingestellt werden können? Wie viele Versicherte bzw. Patient/innen werden das Angebot der eigenen Aktenführung annehmen und in die Weiterverwendung zu Forschungszwecken tatsächlich einwilligen oder ihre Daten dafür spenden? Können insbesondere Personengruppen erreicht werden, die in den bisher verwendbaren Datenbeständen unterrepräsentiert sind? Eine Begleitforschung zu patientengeführten Akten könnte sich derartigen Fragen widmen.

Notwendige technische Datenaufbereitungen für komplexe Analysen

Bisher sind weder die arztgeführten Primärakten noch die diese verwaltenden Informationssysteme für komplexe Datenanalysen oder Data-Mining-Prozesse konzipiert. Dafür müssen zumindest Teile der Patientenakten aufbereitet werden. Die Weiterverwendungsabsicht ist entscheidend für die Aufbereitung: Während für eine bloße interne Archivierung eine wenig strukturierte Ablage ausreicht, ist für eine datenanalytische Weiterverwendung ein strukturiertes Datenmanagementsystem erforderlich. Dafür ist u. a. zu entscheiden, in welchem Umfang neben den Befund- und Diagnosedaten sowie den Behandlungsdokumentationen auch die immer umfangreicheren (Roh-)Daten u. a. von genetischen oder bildgebenden Untersuchungen zu archivieren sind. Da sich mit jeder technischen Weiterentwicklung der datenerfassenden Geräte die Rohdatensätze verändern können, ist ein spezifisches Management der Metadaten erforderlich.

87 derzeitiger Planungsstand ausführlicher in gematik (2021)

Zudem müssen vielfältige Elemente codiert, klassifiziert und in maschinenlesbare Form gebracht werden. Die aus den Rohdaten abgeleiteten Befunde und Diagnosen, die Anweisungen und Dokumentationen zur Behandlung sowie Arztbriefe und Entlassungsdokumente sind in den Primärakten oft als Freitext in nicht maschinenlesbaren Formaten gespeichert. Die Aufbereitung dieser Texte ist bislang personalaufwendig. Zunehmend werden dafür computerlinguistische und sprachtechnologische Verfahren entwickelt, die zumindest Textbestandteile erkennen (Text-Mining) (Schepers et al. 2015, S. 60 ff.). Bisher müssen die damit erzeugten Daten oft noch manuell geprüft und nachbearbeitet werden (Geßner/Dugas 2017, S. 4). Zum Training künstlicher neuronaler Netze mit maschinellen Lernverfahren ist es wichtig, Rohdaten und Befunde spezifisch zu verknüpfen, relevante Datenausschnitte/-elemente/-bereiche zu markieren (Datenlabeling; Kap. 2.2.1). Die kontinuierliche Weiterentwicklung medizinischer Terminologien und Klassifikationen ist dabei eine besondere Herausforderung.

Laut Schepers et al. (2015, S. 183 ff.) setzen vor allem große Krankenhäuser zunehmend auf Data Warehouses als Datenrepositorien (Kap. 2.2.2). Datenschutzverantwortliche und Ethikkommissionen werden bereits in deren Konzeptionierung einbezogen, um sicherzustellen, dass eine mögliche spätere sekundäre Datennutzung im Rahmen geltenden Rechts erfolgt. Die Datenaufbereitung und -migration in diese Repositorien ist mit erheblichem Aufwand verbunden. Die analytischen Potenziale sind groß. Vor allem im öffentlich finanzierten Forschungsbereich gewinnen einrichtungsübergreifende Datennutzungskonzepte an Bedeutung. Derartige Ansätze benötigen sowohl die Einwilligung betroffener Personen als auch Datentreuhandstrukturen, die Datenschutzmaßnahmen sichern und die Zulässigkeit der Analysen prüfen.

4.1.4 Aufbereitete medizinische Datenbestände

Seit Jahren gibt es vielfältige Initiativen, medizinische Daten strukturiert zusammenzuführen, um sie zu unterschiedlichen Zwecken weiterverwenden zu können. Der klassische Ansatz sind medizinische Register, in denen meist auf einzelne Krankheiten bezogene medizinische Behandlungsdatensätze einrichtungsübergreifend zusammengeführt werden. Im Rahmen der Medizininformatik Initiative werden sowohl Forschungs- als auch Behandlungsdatenbestände von Universitätskliniken standardisiert, für sekundäre Analysen aufbereitet und zusammengeführt. Für komplexe Datenanalysen oder Data-Mining-Aktivitäten sind diese Datenbestände hochrelevant, teilweise unmittelbar zur Mustersuche, teilweise mittelbar für Folgeabschätzungen (Kap. 3.3.2).

Medizinische Registervielfalt

Medizinische Register sammeln definierte Datensätze zur Entstehung, Entwicklung und Behandlung einzelner Krankheiten. Aussagen zum Auftreten und zum Verlauf von Krankheiten oder zu den Folgen spezieller Früherkennungs- und Behandlungsmaßnahmen lassen sich empirisch besser absichern, wenn entsprechende Daten einrichtungsübergreifend analysiert werden können. Es gibt zwei methodisch unterschiedliche Formen:

Epidemiologische Register erheben möglichst vollständige krankheitsbezogene Daten für eine Population (Public-Health Perspektive). Sie sind vor allem ein Instrument zur Überwachung der Verbreitung von Krankheiten und zur Planung der Gesundheitsversorgung. Im Fokus stehen einheitlich definierte Zählungen von Neuerkrankungen und von Krankheitsverläufen, um unterschiedliche epidemiologische Kennziffern berechnen zu können (u. a. Inzidenz, Mortalität). Epidemiologische Register gibt es seit langer Zeit zu Infektionskrankheiten sowie zu Krebserkrankungen. Die gegenwärtige tägliche Coronalagebeurteilung beruht auf epidemiologischen Fallzahlmeldungen. Um Vollzähligkeit zu erreichen bedarf es einer gesetzlichen Grundlage.

Klinische Register sammeln umfangreiche Datensätze zu einzelnen Erkrankungsfällen ohne Vollständigkeit anzustreben (medizinische Perspektive auf eine Krankheit). Vorrangige Ziele sind medizinische Beobachtungen zu einzelnen Krankheiten, deren Behandlung, gesundheitsbezogene Folgen sowie Verbesserungen medizinischer Versorgungsprozesse. Je nach Organisation und Struktur der Register können sie auch den Informationsaustausch unterstützen oder zur gezielten Patientenrekrutierung für klinische Studien genutzt werden. Wesentliche Akteure sind medizinische Fachgesellschaften und spezielle medizinische Einrichtungen. Die Führung klinischer Register wird in der Regel nicht als Aufgabe im öffentlichen Interesse aufgefasst (Ausnahme klinische Krebsregister). Für den dauerhaften Betrieb müssen eigenständige Finanzierungsformen und Geschäftsmodelle entwickelt werden. Für die Datenzusammenführung und -verarbeitung gelten die Vorgaben der DSGVO vollumfänglich (Kap. 3.3). Laut Schepers et al. (2015, S. 194 f.) ist es kaum möglich, die Zahl der in Deutschland keinen gesetzlichen Vorgaben unterliegenden klinischen Register genau zu bestimmen. Die Bundesregierung geht von derzeit 270 aktiven medizinischen Fachregistern aus (Bundesregierung 2021c, S. 2). Dazu folgende Beispiele:

Das *TraumaRegister*[88] wurde 1993 von der Arbeitsgruppe Polytrauma in der Deutschen Gesellschaft für Unfallheilkunde (DGU) mit dem Ziel gegründet, die Qualität der Behandlung schwerverletzter Patient/innen zu verbessern. Gegenwärtig beteiligen sich ca. 650 Kliniken vor allem aus Deutschland (ca. 10 %

88 www.traumaregister-dgu.de (10.11.2021)

der teilnehmenden Kliniken kommen aus derzeit 9 weiteren vor allem europäischen Ländern). Sie übermitteln entsprechend der Vorgaben des DGU-Weißbuchs pseudonymisierte einzelfallbezogene Datensätze von Traumapatient/innen, die über den Schockraum der Notaufnahmen eingeliefert werden, an das Register (DGU 2012). Der Gesamtdatenbestand des Registers enthält ca. 0,4 Mio. Behandlungsverläufe (Stand Juni 2021). Neben zeitlichen Entwicklungen und räumlichen Besonderheiten können über den Gesamtbestand häufigste, mittlere und durchschnittliche Behandlungsabläufe sowie einrichtungsspezifische Besonderheiten und Vergleiche zu anderen Einrichtungen ermittelt werden (Benchmarking; Kap. 4.4.2). Teilnehmende Einrichtungen erhalten einrichtungsspezifische Qualitätsberichte. Die Registeraktivitäten werden über Beiträge teilnehmender Kliniken finanziert. Für diese ist es ein Instrument zur Qualitätssicherung und -weiterentwicklung, wozu medizinische Einrichtungen, die Leistungen im Rahmen der GKV erbringen, inzwischen in Deutschland gesetzlich verpflichtet und befugt sind, dazu erforderliche Daten zu verarbeiten (§ 135 ff. i. V. m. § 299 SGB V; ausführlicher in Bundesregierung 2021c). Dennoch sind seit dem Inkrafttreten der DSGVO die Meldungen aus Sorge vor Rechtsverstößen erheblich zurückgegangen.[89]

Gestützt auf die Empfehlungen des EU-Sachverständigenausschuss für seltene Krankheiten hat das Bundesgesundheitsministerium (BMG) die Entwicklung eines ***O**pen-**S**ource-Registersystems für **s**eltene **E**rkrankungen in Europa* (OSSE) gefördert (Muscholl et al. 2016). Damit sollen Patientenorganisationen, behandelnde und/oder forschende Ärzt/innen, medizinische Einrichtungen oder andere Akteure dezentrale, krankheitsspezifische, interoperable Patientenregister aufbauen und mit Patienteneinwilligung definierte pseudonymisierte Minimaldatensätze mit möglichen krankheitsspezifischen Zusatzmerkmalen verwalten. Alle dezentralen Patientenregister werden über eine zentrale Plattform (Register der Einzelregister) orchestriert, bei der lediglich die Metadaten der Einzelregister abgelegt sind und die u. a. Suchanfragen vermittelt. Die jeweiligen dezentralen Einzelregisterbetreiber behalten als verantwortliche Stellen die Datenhoheit, sie entscheiden über das Zustandekommen einer Kooperation mit den Einrichtungen, die eine Suchanfrage gestartet haben. Die ersten zwei Referenzregister wurden 2016 eingeführt. Über die Finanzierung des dauerhaften Betriebs, die Nutzung, die weitere Entwicklung oder mögliche Evaluationen des Konzepts wird über die Homepage des Registersystems nicht berichtet (Stand Anfang 2021).[90]

In Analogie zum OSSE-Ansatz gab es vor einigen Jahren auch Ideen zum Aufbau einer zentralen Registerplattform für alle national existierenden klinischen Register (Stausberg et al. 2014). Realisiert wurden diese Ideen bisher

89 www.dgu-online.de/news-detailansicht/dgu-praesident-uebertriebener-datenschutz-gefaehrdet-menschenleben-in-der-schwerverletztenversorgung.html (10.11.2021)

90 www.osse-register.de (10.11.2021)

nicht. Derzeit gibt es den Vorschlag, die Registerdaten über die Nationale Forschungsdateninfrastruktur (NFDI) zusammenzufügen, zu harmonisieren, nachhaltig zu sichern und der Forschung breiter zugänglich zu machen (Bundesregierung 2020a, S. 13). Es bleibt abzuwarten, ob und (wenn ja) wie sich dieser Vorschlag realisieren lässt.

Krebsregister nehmen in der nationalen Registerlandschaft eine Sonderstellung ein. Nicht nur epidemiologische auch klinische Krebsregister werden inzwischen auf gesetzlicher Grundlage geführt. Die Etablierung der Register war ein jahrzehntelanger Prozess mit vielen Etappen und gesetzlichen Regelungen. Sie begann mit dem Aufbau von epidemiologischen Registern auf Landesebene. Dann wurde zusätzlich zu den Landesregistern ein nationales Zentrum für Krebsregisterdaten (ZfKD) beim Robert Koch Institut eingerichtet, das die epidemiologischen Daten auf Bundesebene zusammengeführt und das diesbezügliche nationale Monitoring und die Berichterstattung organisiert. Zudem wurde die ursprüngliche epidemiologische Meldung einer Krebserkrankung erweitert und die Übermittlung eines umfangreichen klinischen Datensatzes für jede Neuerkrankung vereinbart. Es dauerte Jahre bis die dafür nötigen Verfahren landesrechtlich verankert wurden. Der Start war schleppend (Prognos 2016). Seit 2018 werden für jede an Krebs erkrankte Person definierte medizinische Basisdatensätze in Landesregistern zusammengeführt und zeitlich fortgeschrieben (weitgehende Vollerhebung). Den Aufbau der klinischen Krebsregister hat vor allem die Deutsche Krebshilfe finanziert. Die Kosten für die Erstellung der Einzeldatensätze trägt die GKV (§ 65c SGB V). Bei der Erstevaluierung wurde die Datennutzung insgesamt als kritisch bewertet, im Folgegutachten wurde diesbezüglich ein Fortschritt anerkannt (Prognos 2016, 2018). Um die Datennutzung zu verbessern, wurde inzwischen die Zusammenführung klinischer Krebsregisterdaten beim ZfKD gesetzlich definiert sowie die Ausweitung der Analysemöglichkeiten und die Entwicklung einer Datenplattform anvisiert.[91]

Bisher ist die Zusammenstellung und Übermittlung definierter medizinischer Behandlungsdatensätze in unterschiedliche Register aufgrund fehlender Primäraktenstandards und mangelhafter Interoperabilität zwischen PIS/KIS für medizinische Einrichtungen mit erheblichem Aufwand verbunden. Alle Register haben eigene Treuhandstrukturen aufgebaut. Die Prüfung der eingehenden Datensätze ist auch bei den Registerstellen aufwendig. Teilweise dauert es Jahre, bis die Daten über Registerstellen bereitgestellt werden.[92] Eine automatisierte Datenübertragung von medizinischen Einrichtungen in diverse Register ist nach wie vor mehr Vision als Wirklichkeit (Schepers et al. 2015, S. 183).

91 Gesetz zur Zusammenführung von Krebsregisterdaten

92 Beim ZfKD gibt es derzeit einen Zeitverzug von ca. 3 Jahren (www.krebsdaten.de/; 2.11.2021).

Die nationale Medizininformatik Initiative

Die Medizininformatik Initiative startete 2017 unter Federführung der Universitätskliniken mit Beteiligung von Forschungseinrichtungen, Unternehmen, Krankenkassen und Patientenvertretern. Das Bundesministerium für Bildung und Forschung (BMBF) fördert die Initiative zunächst für den Zeitraum von 2017 bis 2022 mit 180 Mio. Euro. Die an den Universitätskliniken vorhandenen und kontinuierlich hinzukommenden medizinischen Daten sowohl aus Forschungs- als auch aus Behandlungsprozessen sollen zusammengeführt und über die Grenzen einzelner Einrichtungen hinweg für datenbasierte Forschungsaktivitäten genutzt werden können. Nach eigenen Angaben sind die Aktivitätsschwerpunkte zur Entwicklung einer universitätsmedizinischen Dateninfrastruktur:

- *Aufbau von Datenintegrationszentren*: Medizinische Daten aus vielfältigen Informationssystemen und den Archiven sollen strukturiert aufbereitet und klinikweit zusammenführt werden. Externe Vertrauensstellen sichern die reversible Pseudonymisierung ab, Ethikkommissionen prüfen Nutzungsanfragen. Ein Register aller datenanalytischen Projekte soll Transparenz bei der Weiterverwendung sichern.
- *Weiterentwicklungen arztgeführter Patientenaktensysteme*: Die primäre Datenspeicherung soll standardisiert und ein hoher Interoperabilitätsgrad erreicht werden.
- *Entwicklung von Einwilligungsmodellen*: Rechtssichere einheitliche Formulierungen sollen modulare Einwilligungsmöglichkeiten schaffen, die perspektivisch in persönliche Datenmanagementsysteme integriert werden können.

Anhand der Vorgaben der DSGVO zur privilegierten Datenweiterverwendung zu Forschungszwecken (Kap. 3.3.4) wurde ein Datenmanagementmodell sowie ein Mustertext für eine Patienteninformation und für eine breite Patienteneinwilligung erarbeitet und mit der Datenschutzaufsicht abgestimmt (MII 2020). Patient/innen werden um Einwilligung gebeten, dass die jeweils verantwortlichen datenverarbeitenden Stellen (Unikliniken) deren patientenbezogene Forschungs- und Behandlungsdaten langfristig reversibel pseudonymisiert speichern und zu Forschungszwecken verwenden dürfen (eine Kontaktaufnahme soll zu einem späteren Zeitpunkt möglich sein). Ein Treuhandverfahren wird eingerichtet, Daten werden von einer unabhängigen Stelle pseudonymisiert. Eine unabhängige Ethikkommission prüft jeden Forschungsantrag (antragsberechtigt sind öffentlichen Forschungseinrichtungen sowie forschende Unternehmen), bei positivem Votum werden anonymisierte Daten bereitgestellt. Zudem werden Patient/innen zusätzlich um eine Einwilligung gebeten, die bei den Krankenkassen gespeicherten patientenbezogenen Leistungsabrechnungsdaten

des ambulanten Bereichs anfordern zu dürfen, um die eigenen Bestände ergänzen zu können. Es wird auch darauf hingewiesen, dass die verantwortliche datenhaltende Stelle diese Daten nicht verkauft, jedoch bei Weitergabe zu Forschungszwecken eine angemessene Aufwandsentschädigung erheben darf.

Der Aufbau der Dateninfrastruktur wurde mit ersten datenanalytischen Aktivitäten verbunden. Nach 3 Jahren Aufbauarbeit wurde Ende 2020 vor allem auf Erfolge beim Aufbau der medizinischen Dateninfrastruktur hingewiesen. Die Aktivitäten zur intensivierten Datennutzung befanden sich durchgängig noch in der Konzeptphase.[93] Die mit dieser Medizininformatik Initiative verbundenen Erwartungen sind groß, auch wenn keine schnellen datenanalytischen Durchbrüche realisiert werden konnten. Sie ist Bestandteil diverser Strategien der Bundesregierung (Bundesregierung 2018c, 2020a u. 2021a).

4.1.5 Gesamteinschätzung Datenzugänglichkeit

Für die Suche nach Mustern und Strukturen in medizinischen Daten werden derartige Daten von vielen Personen bzw. Patient/innen benötigt, die im nationalen Gesundheitssystem von unterschiedlichen Stellen fragmentiert erhoben und gespeichert werden. Der Datenzugang und die Weiterverwendungsmöglichkeiten werden auf mehrfache Art und Weise begrenzt und lenkt.

Zum Ersten gibt es rechtliche Schutzmechanismen, die die (Grund-)Rechte betroffener Personen sowie teilweise auch die mit der Datenverarbeitung verbundene Leistung schützen. Datenschutz und ärztliche Schweigepflicht verlangen für Datenanalysen entweder zweckgebundene informierte Einwilligungen, die in der Forschungspraxis regelmäßig für breite Analysezwecke eingeholt werden, wobei der Zugang über Datentreuhänder kontrolliert und kanalisiert wird (Fundament klinischer Studien), oder gesetzliche Grundlagen (Fundament der Behandlungsdokumentation), die zwar die Arzt-Patienten-Kommunikation während der Behandlung vereinfacht, jedoch Datenzusammenführungen und Weiterverwendungen begrenzen. Für komplexe Analysen sind erneut entweder gesetzliche Regelungen (bei Aufgaben im öffentlichen Interesse) oder Patienteneinwilligungen erforderlich. An dieser Stelle könnten digitale Einwilligungsmanagementsysteme neue Möglichkeiten der Zusammenführung und Analyse eröffnen. Gesetzliche Datenzugangsverpflichtungen gibt es bisher nicht. Vor allem Daten aus kommerziellen klinischen Studien können Studiensponsoren allein verwerten. Parallel dazu erhalten privatwirtschaftliche Akteure kaum Zugang zu medizinischen Daten, die im Rahmen des öffentlichen Gesundheitssystems generiert wurden.

93 www.medizininformatik-initiative.de (10.11.2021)

Zum Zweiten gibt es technische Barrieren, die die Weiterverwendung vor allem von Behandlungsdaten begrenzen. Die im Rahmen der Behandlung genutzten primären datenverarbeitenden Systeme (PIS/KIS) und deren Datenspeicher sind weder für umfangreiche Datenaufbereitungen noch für komplexe Analysen konzipiert. Aufgrund der geringen Interoperabilität der Primärsysteme ist die Datenüberführung in Data-Mining-geeignete Repositorien bislang aufwendig. Dadurch kommen zum Dritten finanzielle und personelle Barrieren vor allem bei der Aufbereitung von Behandlungsdaten und deren Überführung in spezifische Repositorien hinzu. Diese Aufbereitung wird bisher wesentlich im Rahmen gesetzlicher Behandlungsleistungen solidarisch (über die Krankenversicherungen) oder durch Forschungsprojekte und -initiativen öffentlich finanziert. Diverse krankheitsspezifische Register werden seit Jahren meist mit Unterstützung medizinischer Fachgesellschaften aufgebaut. Im Rahmen der Medizininformatik Initiative führen die nationalen Unikliniken ihre Studien- und Behandlungsdaten in eigenen Repositorien zusammen und visieren zudem erste Data-Mining-Prozesse aber auch eine bessere Interoperabilität bei der Primärdatenhaltung an.

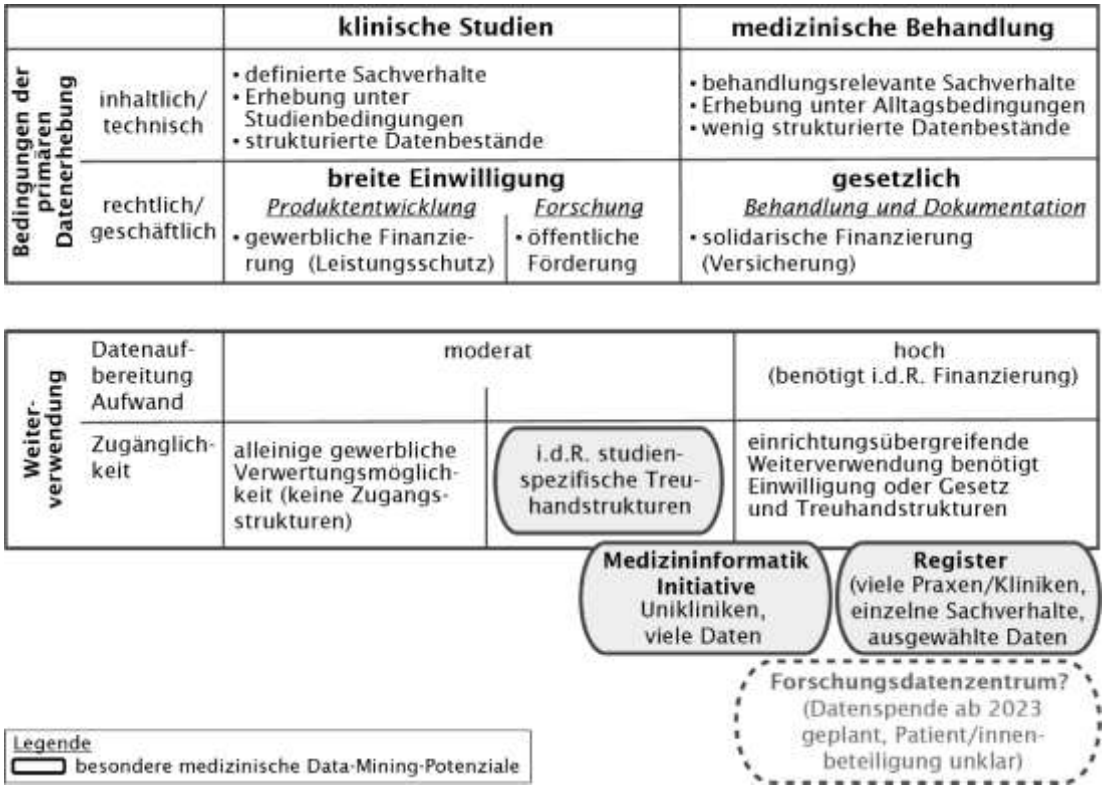
Die dauerhafte Datenbereitstellung mittels Datentreuhandstrukturen benötigt eine eigene Finanzierung. Treuhandverfahren haben sich in der Medizin für die Datenweiterverwendung zu Forschungszwecken weitgehend durchgesetzt. Es gibt jedoch unterschiedliche Ausgestaltungen. Inwiefern die spezifisch definierten, datenanalytischen Ziele der einzelnen Repositorien realisiert werden, kann nur im Einzelfall untersucht und bewertet werden. Eine transparente Darstellung der Analyseanträge und der Resultate wären dafür hilfreich.

Die inhaltlich/technischen und rechtlich/geschäftlichen Strukturen der Entstehung und Weiterverwendung medizinischer Daten sind in Abbildung 4.1 zusammengefasst.

Um eine rechtskonforme Nutzung medizinischer Daten zu befördern, sollte zum einen ein besonderes Augenmerk auf die Weiterentwicklung der Einwilligungsverfahren gelegt werden. Bisher werden rechtssichere Einwilligungen in medizinischen Kontexten schriftlich eingeholt. Diesbezügliche Weiterentwicklungen sollten selektive oder dynamische Einwilligungen bis hin zu in Personal Managements Systemen ermöglichen. Die patientengeführten elektronischen Patientenakten könnten die dafür notwendige Kommunikationsplattform werden. Zum anderen sollte bei der Weiterentwicklung der Treuhandstrukturen die Datennutzung stärker in den Blick genommen werden. Dazu sollten die Prüfprozesse der Nutzungsanträge vereinheitlicht und verkürzt sowie die Transparenz bei der Weiterverwendung durch Antrags-, Ergebnis- oder Publikationsregister erhöht werden. Die Bereitstellung von datenanalytischen Werkzeugen ggf. auch von Serviceteams, die Nutzende bei ihren Aktivitäten unterstützen, könnten die Datennutzung ebenfalls erleichtern. Eine vertiefte Auseinanderset-

zung mit den Treuhandstrukturen in der Medizin und den gesammelten Erfahrungen, könnte auch für andere Bereiche hilfreich sein, die personenbeziehbare Daten über Forschungsdatenzentren und Dateninfrastrukturen bereitstellen (Kap. 3.3.3).

Abb. 4.1 Erhebung medizinischer Datenbestände und deren Zugänglichkeit im Überblick



Eigene Darstellung

4.2 Medizinprodukte zur Generierung und Analyse medizinischer Daten

Das Medizinprodukterecht rahmt medizinische Data-Mining-Prozesse in besonderem Maße: Zum einen sind die Messgeräte, die Daten zur Diagnose und Behandlung von Krankheiten erfassen, Medizinprodukte. Zum anderen fallen aus Data-Mining resultierende algorithmische Systeme, sofern sie zu digitalen Anwendungen oder Softwarebestandteilen weiterentwickelt werden und krankheitsbezogene Informationen über einzelne Personen liefern, unter das Medizinprodukterecht: Produkte einschließlich Software, die u. a. Krankheiten oder Behinderungen diagnostizieren, überwachen, prognostizieren oder einen physi-

ologischen Zustand untersuchen sollen, sind aus rechtlicher Sicht *Medizinprodukte* (Art. 2 Nr. 1 Verordnung (EU) 2017/745) bzw. *In-vitro-Diagnostika*, wenn sie aus dem menschlichen Körper stammende Proben analysieren (Art. 2 Nr. 2 Verordnung (EU) 2017/746).[94]

Das Medizinprodukterecht[95] definiert ein spezifisches *Qualitätsmanagementsystem* mit unterschiedlichen Elementen je nach Gesundheitsschädigungspotenzial eines Produktes. Es soll eine qualitativ hochwertige Datenerfassung und -analytik sichern. Hersteller tragen die Verantwortung dafür. Sie müssen die Sicherheit und Leistungsfähigkeit ihres Produktes prüfen (Entwicklungsphase), bevor es zertifiziert (Markteintritt) und umfangreich eingesetzt werden kann (Anwendungsphase).

Medizinprodukte werden entsprechend der jeweils möglichen gesundheitsbezogenen Folgen für betroffene Patient/innen vier verschiedenen Risikoklassen zugeordnet (dazu und im Folgenden Geißelmann 2018; Lücker 2018). Mit Inkrafttreten der europäischen Medizinprodukteverordnungen gehören in Klasse I und a. einfache Mess- und Datenspeicherprodukte (z. B. Blutdruckmesser oder krankheitsbezogene Tagebücher), sofern damit lediglich krankheitsbezogene Vitalwerte und Zustände erfasst und aufgezeichnet werden, um anwendende Personen (sowohl Ärzt/innen als auch Patient/innen) zu informieren. Schon wenn Durchschnittswerte, Trends oder Klassifikationswerte (medizinische Scores) berechnet werden, gelten sie als aktive diagnostische Medizinprodukte, die in der Regel in die Risikoklasse IIa fallen. Abhängig von möglichen gesundheitlichen Folgen können sie auch in die Risikoklasse IIb (z. B. bei der Überwachung von Herzfunktionen) oder die höchste Klasse III (z. B. Berechnung der Dosierung hochwirksamer Medikamente) fallen, vor allem wenn sie folgenschwere ärztliche Tätigkeiten übernehmen und automatisieren. Informationssysteme, die lediglich Daten speichern, präsentieren und für Versicherungs- und Abrechnungszwecke verarbeiten sowie Software ohne medizinische Zweckbestimmung (keine Anwendung im Kontext von Krankheiten) gelten nicht als Medizinprodukte. Es gibt fließende Übergänge und einen zunehmenden Graubereich zwischen Software mit und ohne medizinische Zweckbestimmung, insbesondere im wachsenden Markt der Gesundheits-Apps, die ein immer größeres Spektrum an Vitalwerten erfassen (ausführlich z. B. TAB 2018).

---

94 Verordnung (EU) 2017/745 über Medizinprodukte, zur Änderung der Richtlinie 2001/83/EG, der Verordnung (EG) Nr. 178/2002 und der Verordnung (EG) Nr. 1223/2009 und zur Aufhebung der Richtlinien 90/385/EWG und 93/42/EWG; sie ist am 26. Mai 2021 in Kraft getreten. Verordnung (EU) 2017/746 über In-vitro-Diagnostika und zur Aufhebung der Richtlinie 98/79/EG und des Beschlusses 2010/227/; diese soll am 26. Mai 2022 in Kraft treten. Nachfolgend wird vereinfachend allgemein von Medizinprodukten und vom Medizinprodukterecht gesprochen, darin sind auch In-vitro-Diagnostika eingeschlossen.

95 Überblick über alle zum Medizinprodukterecht gehörenden europäischen und nationalen Rechtsnormen z. B. unter www.pei.de/DE/service/rechtliches/medizinprodukterecht/medizinprodukterecht-node.html (10.11.2021)

Unter Achtung des geltenden Rechts und der Berücksichtigung der Rechtsprechung in Einzelfällen hat ein Hersteller vor allem bei datenerhebenden und -verarbeitenden (Software-)Produkten niedriger Risikoklassen einen gewissen Entscheidungsspielraum, diese als Gesundheits-App oder als Medizinprodukt zu vermarkten. Im ersten Gesundheitsmarkt sind nur Medizinprodukte verordnungs- und erstattungsfähig.

4.2.1 Prüfung und Bewertung der Sicherheit, Leistung und Gesundheitseffekte

Medizinprodukte dürfen nur mit CE-Kennzeichnung[96] und nach behördlicher Anzeige (in Deutschland beim BfArM) in den Verkehr gebracht werden. Hersteller müssen dafür ein Konformitätsbewertungsverfahren durchführen und darlegen, dass ihr Produkt eine definierte Leistung erbringt (z. B. einen Vitalwert korrekt misst, einen Grenzwert richtig berechnet, ein Muster richtig erkennt) und grundlegende Anforderungen an Qualität, Sicherheit und Unbedenklichkeit erfüllt werden (einschließlich DSGVO-Standards für personenbezogene Daten besonderer Kategorie; Kap. 3.3.2; Kasten 3.6). Die Leistung für einen vorgesehenen Verwendungszweck wird in normierten Verfahren[97] anhand klinischer (medizinischer) Daten nachgewiesen. Relevante Kriterien zur Leistungsbewertung sind vor allem die Sensitivität und Spezifität eines Verfahrens (Kap. 2.3.3). Neue Datenerhebungs- und Datenanalysetechniken werden nach Möglichkeit mit bereits etablierten ähnlichen Verfahren verglichen. Schwieriger ist die Bewertung neuartiger Verfahren (z. B. Erfassung neuer Biomarker), zu denen es kaum Vergleichsverfahren und Standards gibt (ausführlicher in den Anwendungsbeispielen in Kap. 4.3.3).

Die Anforderungen an die Leistungsbewertung steigen mit der Risikoklasse. Bei Medizinprodukten der Risikoklasse I führt der Hersteller die Leistungsbewertung eigenverantwortlich durch und hält entsprechende technische Dokumentationen vor, falls staatlich benannte Stellen als Prüfinstanzen (z. B. TÜV) Einsicht fordern. Ab Risikoklasse IIa werden diese Stellen an der Leistungsbewertung beteiligt. Sie übernehmen mehr Prüfungs- und Zertifizierungsaufgaben. Bei Produkten der Risikoklassen IIb und III müssen mit Inkrafttreten

96 Mit dem CE-Kennzeichen erklärt der Hersteller, dass das Produkt den geltenden Anforderungen genügt gemäß Verordnung (EG) Nr. 765/2008 über die Vorschriften für die Akkreditierung und Marktüberwachung im Zusammenhang mit der Vermarktung von Produkten und zur Aufhebung der Verordnung (EWG) Nr. 339/93, bei Medizinprodukten, dass die Vorgaben der Verordnungen (EU) 2017/745 und 2017/746 eingehalten werden.

97 DIN EN ISO 14155 (Klinische Prüfung von Medizinprodukten an Menschen – Gute klinische Praxis [ISO 14155:2020]; Deutsche Fassung EN ISO 14155:2020) und DIN EN ISO 13485 (Medizinprodukte – Qualitätsmanagementsysteme – Anforderungen für regulatorische Zwecke [ISO 13485:2016]; Deutsche Fassung EN ISO 13485:2016 + AC:2018 + A11:2021)

der europäischen Medizinprodukteverordnungen Sicherheit und Leistung anhand von klinischen (Interventions-)Studien (Kap. 4.1.3) belegt werden. Diese klinischen Studien im Rahmen der Produktentwicklung sind genehmigungspflichtig und werden registriert (in Deutschland durch das BfArM). Dafür ist auch ein positives Votum der jeweils zuständigen Ethikkommission erforderlich.

Benannte Stellen prüfen die Angemessenheit der jeweiligen Verfahren sowie Sicherheits- und Leistungsbelege anhand definierter Prüfkriterien, bewerten bei Medizinprodukten hoher Risikoklassen die gesundheitsbezogenen Effekte und zertifizieren das jeweilige Medizinprodukt gegebenenfalls. Sowohl Medizinprodukthersteller als auch benannte Stellen werden akkreditiert, sind mittels Kennziffer eindeutig identifiziert und unterliegen einem Qualitätsmonitoring.

Mit dem Inkrafttreten des überarbeiteten Medizinprodukterechts werden auch etliche Elemente der *Qualitätsmanagementsysteme* für Medizinprodukte ausgebaut, die darauf abzielen, gesundheitsbezogene Risiken während der Anwendung zu minimieren. Dazu gehört u. a. die herstellerseitige Produktregistrierung auf nationaler Ebene im »Deutschen Medizinprodukte-Informations- und Datenbanksystem« (DMIDS).[98] Für die Registrierung neuer Medizinprodukte werden inzwischen europaweit eindeutige Produktkennzeichnungen (Unique Device Identification – UDI[99]) verwendet. Weitere Elemente des Qualitätsmanagementsystems in der Anwendungsphase sind produktspezifisch definierte regelmäßige technische Prüfungen (z. B. Kalibrierungen, Ringversuche), Anleitungen und Schulungen zur Bedienung, ein Beobachtungs- und Meldesystem für mögliche Auffälligkeiten und Risiken während der Anwendung (Vigilanz), klinische Nachbeobachtungen, Hersteller- und Verfahrensüberprüfungen, Verwendungskorrekturen bis hin zu Rückrufen.

Die CE-Kennzeichnung macht ein Medizinprodukt im europäischen Binnenmarkt verkehrsfähig. Insbesondere bei neuartigen aufwendigen und damit kostenintensiven Verfahren stellen sich Fragen zur Integration in medizinische Versorgung und zur Haftung im Schadensfall.

4.2.2 Integration in die medizinische Versorgung

Nicht alles, was technische Geräte inzwischen mit hoher Präzision erfassen und Algorithmen errechnen oder zuordnen können und was unter Studienbedingungen positive Gesundheitseffekte erreicht und zertifiziert wird, generiert in der Anwendung auch einen tatsächlichen gesundheitsbezogenen (Zusatz-)Nutzen. Ärzt/innen sollen mit ihren Fachkenntnissen u. a. die datenanalytischen Verfahren und Medizinprodukte einsetzen, die sachlich gerechtfertigt und medizinisch

98 www.bfarm.de/DE/Medizinprodukte/Portale/DMIDS/_node.html (10.11.2021)
99 europaweit eindeutige Nummer für Medizinprodukte

vertretbar sind. Ein wichtiger Bezugspunkt dafür sind Behandlungsleitlinien, in denen medizinische Fachgesellschaften den Wissensstand zur Diagnose und Behandlung einzelner Krankheitsbilder zusammenfassen, den Einsatz datenanalytischer Verfahren und Medizinprodukte fachlich bewerten und ggf. empfehlen. Ein zweiter relevanter Bezugspunkt sind die Leistungskataloge der Krankenkassen, die die Einsatzkosten neuer datenanalytischer Verfahren ggf. tragen. Die GKV verlangt neben den zertifizierungsrelevanten Sicherheits- und Leistungsnachweisen zunehmend Belege für einen (Zusatz-)Nutzen, bevor der Einsatz neuer Medizinprodukte als Regelleistung erstattet wird. Derartige Nachweise sind sowohl für Hersteller als auch für Prüfinstanzen eine zusätzliche Herausforderung, denn Parameter, anhand derer der Nutzen definiert wird (klinische oder auch lebensstilbezogene und wirtschaftliche Endpunkte), Studiendesigns und Nachweisverfahren müssen oftmals noch weiterentwickelt und entsprechende methodische Standards vereinbart werden (ausführlicher z. B. svr Gesundheit 2014, S. 160 ff.). Bei neuartigen datenanalytischen Verfahren dauert es mitunter Jahre, um diesen (Zusatz-)Nutzen anhand klinischer Studien nachzuweisen, zu prüfen und zu bewerten (ausführlicher in den Anwendungsbeispielen in Kap. 4.3.3). Mit der Anerkennung eines solchen (Zusatz-)Nutzens sind die Voraussetzungen geschaffen, um ein Mess- und Analyseverfahren und entsprechende Medizinprodukte zum einen in krankheitsspezifischen Behandlungsleitlinien der medizinischen Fachgesellschaften zu empfehlen und diese zum anderen in die Leistungs- und Abrechnungskataloge der GKV zu aufzunehmen.

Dieser Weg in das Leistungsportfolio der GKV verläuft im ambulanten und stationären Sektor in unterschiedlichen Bahnen mit vielen Zwischenstationen, Verzweigungen und mit unterschiedlichen beteiligten Institutionen: Im *ambulanten Bereich* können Ärzt/innen standardisiert verschreiben und abrechnen, was explizit erlaubt ist. Dieses Erlaubnisverfahren wurde für digitale Medizinprodukte der Risikoklassen I und IIa durch das Digitale Versorgungs-Gesetz (DVG)[100] mit dem definierten Fast-Track-Verfahren für digitale Gesundheitsanwendungen [DiGA] beschleunigt. Mit den für die CE-Kennzeichnung notwendigen Sicherheits- und Leistungsnachweisen und einem Hinweis auf einen positiven Versorgungseffekt beantragen Hersteller die Aufnahme ihrer Gesundheitsanwendung in ein spezifisches DiGA-Verzeichnis. Das BfArM prüft die Anträge innerhalb weniger Monate und nimmt die Gesundheits-App ggf. in das Verzeichnis auf. Damit gehört die App zunächst vorläufig zum GKV-Leistungsportfolio, Ärzt/innen können sie verschreiben, gesetzliche Krankenkassen tragen die Kosten für ihre Versicherten und erstatten die zunächst vom Hersteller

100 Gesetz für eine bessere Versorgung durch Digitalisierung und Innovation (Digitale-Versorgung-Gesetz – DVG), BGBl. I Nr. 49 vom 18.12.2019, S. 2562–2585

festgelegten Preise. Innerhalb eines Jahres müssen Hersteller den positiven Gesundheitseffekt ihrer App nachweisen, um dauerhaft im DiGA-Verzeichnis gelistet zu werden. Der finale Preis wird dann zwischen dem Hersteller und dem GKV-Spitzenverband verhandelt. Der ambulante Einsatz risikoreicher Medizinprodukte wird durch Richtlinien des Gemeinsamen Bundesausschusses (G-BA), dem obersten Entscheidungsgremium der GKV-Selbstverwaltung erlaubt. Ein ggf. veränderter Betreuungsaufwand bei Ärzt/innen (z. B. wenn sie Vitalwerte und daraus abgeleitete Hinweise auf gesundheitliche Veränderungen bei ihren Patient/innen kontinuierlicher prüfen und in der Behandlung berücksichtigen müssen) muss zudem in deren Abrechnungskatalogen[101] verankert werden. Jenseits dieses Regeleinsatzes kann eine Kostenübernahme für einen Einsatz neuartiger Verfahren mit medizinischer Begründung auch individuell bei der jeweiligen Krankenkasse beantragt und bei Bewilligung auch abgerechnet werden. Zudem können Ärzt/innen neue datenanalytische Verfahren, auch als individuelle Gesundheitsleistung (IGeL) anbieten, Patient/innen tragen anfallende Kosten sowohl für das jeweilige Medizinprodukt als auch für den ärztlichen Aufwand selbst (zweiter Gesundheitsmarkt).

Im *stationären Bereich* haben medizinische Einrichtungen vor allem beim Einsatz risikoreicher Medizinprodukte einen größeren Handlungsspielraum. Ärzt/innen können im Rahmen der gewissenhaften Berufsausübung eigenverantwortlich vielfältige datenanalytische Verfahren einsetzen, die dem anerkannten Stand des Wissens entsprechen, sofern sie nicht durch G-BA-Richtlinien explizit ausgeschlossen werden (Anwendungsbeispiel in Kap. 4.3.3). Bei kostenintensiven neuen Untersuchungs- und Behandlungsmethoden, wird deren schrittweise Integration in die stationären Abrechnungskataloge (G-DRG-System; Kap. 5.2.1) wichtiger.

4.2.3 Haftung und Schadensausgleich

Bei jeglichen Medizinprodukten einschließlich solchen zur Datenerhebung und -analyse (algorithmische Systeme zur Unterstützung medizinischer Entscheidungen als Ergebnis von Data-Mining-Prozessen) wird die Produkthaftung in der klinischen Entwicklungsphase vor der Zertifizierung von der in der Anwendungsphase unterschieden (dazu und im Folgenden Deutscher Ethikrat 2017, S. 108). In der Entwicklungsphase, in der Sicherheit und Leistung jeglicher Medizinprodukte nachgewiesen und geprüft werden (bei hohen Risikoklassen mittels klinischer Studien), haften algorithmenentwickelnde Stellen und erprobende Ärzt/innen gemeinsam bei jeglichen Gesundheitsschäden von Proband/

101 Ärzt/innen rechnen ambulant erbrachte medizinische Leistungen im Rahmen der GKV regulär anhand von Kennziffern und Katalogen des Einheitlichem Bewertungsmaßstabs (EBM) bei den jeweiligen gesetzlichen Krankenkassen ab. Andere Leistungen stellen sie anhand der Gebührenordnung Ärzte (GO-Ä) in Rechnung.

innen, egal wer oder was den Schaden verursacht hat und müssen sich diesbezüglich angemessen versichern (Gefährdungshaftung, Haftpflichtversicherung). Geschädigte Personen müssen nicht nachweisen, dass ein Produktfehler oder eine Pflichtverletzung algorithmenentwickelnder Stellen oder beteiligter Ärzt/innen den Schaden verursacht hat.

Beim regulären Einsatz zertifizierter Medizinprodukte gelten für herstellende bzw. inverkehrbringende Stellen die Vorgaben der allgemeinen Produkthaftung laut Produkthaftungsgesetz (das Medizinprodukterecht zielt mit dem Ausbau der Qualitätsmanagementsysteme auf die Gewährleistung einer hohen Produktsicherheit, reguliert jedoch keine Haftungsfragen beim regulären Einsatz). Herstellende/inverkehrbringende Stellen sind für die Zuverlässigkeit ihrer Produkte im Allgemeinen bzw. ihrer algorithmischen Systeme zur Unterstützung medizinischer Entscheidungen verantwortlich. Sie haften, wenn sie ihre Pflichten rechtswidrig und schuldhaft verletzen oder fehlerhafte Produkte in den Verkehr bringen (z. B., wenn sie wissentlich fehlerhafte Algorithmen in eine Software implementieren) und dies ursächlich für einen tatsächlich aufgetretenen Schaden ist (§ 823 BGB, § 1 ProdHaftG). Fehlerhaft ist ein Medizinprodukt, wenn es nicht die Sicherheit bietet, die entsprechend der Leistungsbewertung erwartet werden kann (wenn z. B. die Sensitivität eines Verfahrens in der Anwendung schlechter ist, als unter Studienbedingungen errechnet). Herstellende/inverkehrbringende Stellen sind für die kontinuierliche Sicherung der Qualität ihrer Produkte in der gesamten Nutzungsphase verantwortlich und müssen dafür das skizzierte Qualitätsmanagementsystem aufbauen und einhalten. Sie haften nicht, wenn durch den ordnungsgemäßen Einsatz eines Medizinproduktes Fehler und in der Folge Schäden auftreten (d. h., sie haften nicht, wenn z. B. eine einzelne Situation falsch positiv oder negativ bewertet wird, dies aber im Rahmen der geforderten produktspezifischen Leistungsfähigkeit bzw. Sensitivität und Spezifität liegt). Im Schadensfall (ausschließlich gesundheitliche und materielle) muss die geschädigte Person nachweisen, dass eine Pflichtverletzung des Algorithmenbereitstellers oder ein Softwarefehler ursächlich für einen Schadenseintritt waren. Eine bloße Gesundheitsgefährdung reicht nicht aus (Verschuldenshaftung).

Da Medizinprodukte im Allgemeinen und algorithmische Systeme zur Unterstützung medizinischer Entscheidungen im Besonderen derzeit wesentlich im Rahmen der Behandlung in der Verantwortung von Ärzt/innen eingesetzt werden, kommt zur Klärung der Herstellerhaftung auch die Arzthaftung (betrifft nur die sorgfältige Vorgehensweise einschließlich Dokumentation, nicht aber einen Behandlungserfolg/Gesundheitsnutzen) verbunden mit Datenschutz- und Schweigepflichten in Bezug auf den Einsatz des algorithmischen Systems bei anderen Patient/innen hinzu. Geschädigte Personen müssen den Fehlernachweis erbringen, können wegen der Datenschutz- und Schweigepflichten jedoch kaum

andere möglicherweise ebenfalls Betroffene ausfindig machen, um nachzuweisen, dass Fehlerquoten nicht im Rahmen der nachgewiesenen Leistungsfähigkeit liegen und das eingesetzte algorithmische Systeme ursächlich für bestimmte Schäden waren. In dieser Konstellation ist die Produktüberwachung während der Anwendung von besonderer Bedeutung (Vigilanz). Entsprechende Monitoringstrukturen werden derzeit nach dem Vorbild der Arzneimittelüberwachung auch für Medizinprodukte ausgebaut.

Auch wenn qualitätsgeprüfte zertifizierte algorithmische Systeme im Rahmen der Behandlung gewissenhaft eingesetzt werden, können deren Ergebnisse falsch sein (weil deren Sensitivität und Spezifität in der Regel unter 100 % liegen und Ärzt/innen fehlerhafte Resultate nicht immer erkennen). In dieser Konstellation haften weder Hersteller noch behandelnde Ärzt/innen. Geschädigte haben zwar grundsätzlich Anspruch auf weitere Behandlungen. Damit lassen sich jedoch nicht immer alle Folgen ausgleichen, vor allem dann nicht, wenn dauerhafte gesundheitliche Beeinträchtigungen verbleiben. In der Vergangenheit wurden bereits für einige spezielle Schadenssituationen zusätzliche Regelungen getroffen und Maßnahmen vereinbart, um Folgeschäden zumindest abzumildern:

- Bei Gesundheitsschäden infolge staatlich empfohlener Impfungen erhalten Betroffene auf Antrag Unterstützung entsprechend dem Bundesversorgungsgesetz[102] (§ 60 Infektionsschutzgesetz).
- Der Deutsche Bundestag hat zudem mehrfach politische Weichen für spezielle Hilfsfonds gestellt. Beispiele sind die Conterganstiftung für behinderte Menschen, die Stiftung humanitäre Hilfe für durch Blutprodukte HIV-infizierte Personen oder der Hilfsfonds für Dopingopfer.

In Österreich und Frankreich gibt es einen allgemeinen Medizinhilfefonds, der gesundheitliche Folgeschäden abmildert, die trotz sorgfältiger Arbeitsweise von behandelnden Ärzt/innen und von Arzneimittel- oder Medizinprodukteherstellern entstanden. In Deutschland wird ein solcher Fonds unter der Bezeichnung Patientenentschädigungs- und Härtefallfonds seit Jahren diskutiert, ohne dass dafür eine politische Mehrheit gefunden werden konnte (ausführlicher z. B. in WD 2016). Folgeschäden durch falsch positive oder falsch negative Ergebnisse medizinischer Tests oder algorithmischer Systeme, die nicht durch zusätzliche Behandlungen ausgeglichen werden können, tragen nach derzeitiger Rechtslage betroffene Patient/innen allein.

Die Regelungen zur Produkt- und Arzthaftung werden vor allem bei risikoreichen medizinischen Ansätzen immer wieder diskutiert. Auch in Bezug auf datentrainierte, medizinische Entscheidungs(unterstützungs)systeme sollte die Angemessenheit der existierenden Regelungen, diskutiert werden. Vor allem

102 Gesetz über die Versorgung der Opfer des Krieges (Bundesversorgungsgesetz – BVG)

bei Systemen, die kontinuierlich lernende Verfahren einsetzen, durch die sich die Sensitivität und Spezifität auch während der Anwendung verändern kann, werden die etablierten Verfahren zur Produkthaftung, bei denen Geschädigte Sorgfaltspflichtverletzungen nachweisen müssen, als kritisch bewertet. Welche Maßnahmen bei datentrainierten algorithmischen Systemen zur Unterstützung medizinischer Entscheidungen am besten geeignet sind, Risiken zu minimieren sowie Schäden ggf. auszugleichen, sollte eingehender untersucht werden. Dabei könnten die österreichischen und französischen Diskussionen und Erfahrungen in Bezug auf deren Medizinhilfefonds zusätzliche Anregungen geben.

4.3 Data-Mining-Anwendungsbeispiele

Unter Berücksichtigung von Leitlinien und Leistungskatalogen der Kostenträger entscheiden Ärzt/innen bei der Behandlung ihrer Patient/innen eigenständig, wie Symptome zu interpretieren, welche Untersuchungen situativ nötig und wie deren Ergebnisse zu bewerten sind, welche gesundheitliche Entwicklung erwartbar ist und welche therapeutischen Optionen möglich sind, um dann mit den Betroffenen die situativ beste Behandlungsoption auszuwählen und zu realisieren. Big-Data- bzw. KI-Protagonisten gehen davon aus, dass vielfältige Algorithmen und Spezialsoftware Ärzt/innen nicht nur bei Einzelaktivitäten wie der Bildbefundung, sondern auch bei komplexen Aufgaben zur Diagnose, Prognose oder gar der Behandlungsplanung zukünftig zumindest unterstützen können. Grundsätzlich können vielfältige ärztliche Tätigkeiten auch als mathematische Aufgabe formuliert werden, um sie mit unterschiedlichen mathematisch-statistischen Verfahren datenbasiert zu lösen (Kap. 2.3.1):

- Klassifikation patientenbezogener Situationen und Sachverhalte (Scoring);
- Suche nach Auffälligkeiten und Strukturen in Daten (Mustererkennung);
- Vorhersagen zum Krankheitsverlauf (prädiktive Modelle) sowie
- Zusammenfassung von Symptomen, Befunden, Diagnosen, Prognosen und Präferenzen sowie Optimierung von Behandlungsabläufen (medizinische Assistenzsysteme).

Anhand unterschiedlicher Anwendungsspeispiele soll nachfolgend ein Einblick gegeben werden in bereits erzielte Erfolge von Data-Mining-Ansätzen sowie methodische Grenzen und spezifische Herausforderungen, die sich ergeben, wenn algorithmische Entscheidungs(unterstützungs)systeme in Behandlungsprozesse integriert werden sollen.

Die Abläufe zur Früherkennung, Diagnostik und Therapie von Brustkrebs sollen nachfolgend einen anwendungsorientierten Rahmen bilden, in dem das Spektrum von Data-Mining-Ansätzen in Behandlungsprozessen umrissen werden soll. Es wird skizziert, welche Datenbestände für Data-Mining-Prozesse bereits herangezogen werden, wie Algorithmen in Behandlungsabläufen bereits

eingesetzt werden und welche Herausforderungen die Integration neuer datenbasierter Verfahren in nationale Früherkennungs- und Behandlungsprogramme mit sich bringen.

4.3.1 Risikoklassifikation und medizinisches Scoring

Brustkrebs ist die häufigste Krebserkrankung bei Frauen in Deutschland.[103]. Das von nationalen Fachgesellschaften diesbezüglich empfohlene Vorgehen für Früherkennung, Diagnostik, Therapie und Nachsorge wird in einer spezifischen interdisziplinären Leitlinie zusammenfassend dargestellt und regelmäßig aktualisiert (DKG et al. 2021). Das nationale Programm zur Brustkrebsfrüherkennung umfasst mehrere Maßnahmen, die gegenwärtig an folgenden Hauptrisikofaktoren ausgerichtet sind:

- *Geschlecht*: Brustkrebs tritt hauptsächlich bei Frauen auf (aber nicht ausschließlich).
- *Lebensalter*: Die Wahrscheinlichkeit an Brustkrebs zu erkranken, steigt bis zum 70. Lebensjahr kontinuierlich. Sie ist bei Frauen ohne familiäre Belastung zwischen dem 60. und 70. Lebensjahr am höchsten.
- *Familiäre Belastung/genetische Prädisposition*: 5 bis 10 % der Neuerkrankungen treten bei Mitgliedern brustkrebsbelasteter Familien auf. Bei 25 bis 50 % dieser Neuerkrankten werden bestimmte Mutationen in mindestens einem von zwei Breast-Cancer-Genen (BRCA) gefunden.

Etliche Risikofaktoren wie z. B. die Eingrenzung der Lebensphase, in der die Erkrankungswahrscheinlichkeit deutlich erhöht ist, können nur empirisch fundiert werden, wenn Daten von vielen Personen zusammengeführt und retrospektiv analysiert werden. Die Datenbasis dafür waren zunächst umfangreiche Beobachtungsstudien, inzwischen können zunehmend auch Daten von Krebsregistern genutzt werden (Kap. 4.1.3). Für die Analyse wurden vor allem klassische symbolische Verfahren eingesetzt (Kap. 2.3.2). Relevant sind die Schlüsse, die aus derartigen Analysen gezogen werden. Auf der gesundheitssystemischen Ebene wurde u. a. ein nationales Mammografiescreeningprogramm für alle Frauen zwischen dem 50. und dem 70. Lebensjahr aufgelegt, um Brustkrebs möglichst in frühen Entwicklungsphasen zu erkennen (dann sind die Heilungschancen am größten). Zudem gibt es ein intensiviertes medizinisches Betreuungsprogramm für Frauen mit familiärer Belastung. Nach ihnen wird mit einem gestuften Filterverfahren gezielt gesucht. Die Entwicklung dieser Filterverfahren beruht auf unterschiedlichen datenanalytischen Ansätzen.

103 Die Informationen zu Brustkrebs entstammen wesentlich dem Krebsinformationsdienst (www.krebsinformationsdienst.de/tumorarten/brustkrebs/index.php; 10.11.2021).

Bereits vor der Jahrtausendwende wurden anhand der Daten umfangreicher Beobachtungsstudien mit retrospektiven Analysen unterschiedliche einfachste familiäre Alltagskriterien[104] ermittelt, die jeweils mit einer mehr als 10%igen empirischen Wahrscheinlichkeit einhergehen, dass bei einer Frau eine Genmutation vererbt wurde, die ein erhöhtes Erkrankungsrisiko mit sich bringt.

Aus methodischer Sicht könnte man die datenbasierte Ermittlung der Kriterien als Data-Mining bezeichnen. Das Ergebnis sind trivial einfache Klassifikationsregeln, mit denen ein an sich komplexer Sachverhalt (familiäre Brustkrebsbelastung) durch einen binären Risikowert vereinfacht dargestellt wird. Natürlich haben derartig grobe Vereinfachungen methodische Schwächen (z. B. keine Differenzierung zwischen großen und kleinen Familien, Grenzwertfestlegungen, die eine Risikoberatung indizieren). Die Stärke derartig einfacher Checklisten liegt in der transparenten Darstellung der Regeln und der einfachen Anwendung im Rahmen von Ärzt/innen-Patient/innen-Gesprächen. Derartige Listen sind einfachste Hilfsmittel für Ärzt/innen. Deren Einsatz wird allgemein dem ärztlichen Handeln im Rahmen der Anamnese zugerechnet. Eine digitale Anwendung ist dafür nicht erforderlich. Die bloße Checkliste ist rechtlich kein Medizinprodukt (Kap. 4.2).

Bei einem wahrscheinlich erhöhten Erkrankungsrisiko sollte eine spezifischere Risikoberatung in einem Zentrum für familiären Brust- und Eierstockkrebs empfohlen werden. Diese an etlichen Unikliniken angesiedelten Zentren bieten zunächst eine genaue prädiktive Basisdiagnostik an, um das jeweilige Erkrankungsrisiko spezifischer zu bestimmen. Dazu wird eingangs die familiäre Erkrankungssituation anhand eines Tableaus mit 26 Kriterien differenzierter abbildet und jedes Kriterium spezifisch gewichtet (da sie die erbliche Belastung unterschiedlich stark indizieren). Das Ergebnis ist ein medizinischer Risiko-Score, mit dem die Erkrankungswahrscheinlichkeit für einzelne Patient/innen genauer bewertet werden kann (DKG 2016). Die Datenbasis für die Erstellung ausdifferenzierter Kriterienkataloge bilden inzwischen Krebsregister (Kap. 4.1.4) sowie die Forschungsdatenbanken der Zentren für familiären Brust- und Eierstockkrebs. Wenn derartige Kataloge und Risiko-Score-Berechnungen digitalisiert werden, wird aus rechtlicher Sicht die Stufe zum Medizinprodukt (mit geringer Risikoklasse) überschritten. Ein Konformitätsbewertungsverfahren und eine Zertifizierung werden erforderlich (Kap. 4.2).

Dieses Einstiegsbeispiel soll verdeutlichen, dass auch vergleichsweise einfache Data-Mining-Prozesse, die auf klassischen statistischen Verfahren aufbauen, zu hilfreichen Werkzeugen für die medizinische Praxis führen können.

---

104 In einer Linie der Familie erkrankte(n) mindestens (DKG et al. 2021, S. 55):
3 Frauen an Brustkrebs; 2 Frauen an Brustkrebs (davon eine unter 50 Jahre); 2 Frauen an Eierstockkrebs; 1 Frau an Brustkrebs und 1 Frau an Eierstockkrebs; 1 Frau an Brust- und Eierstockkrebs; 1 Frau unter 36 Jahren an Brustkrebs; 1 Frau unter 50 Jahren an bilateralem Brustkrebs oder 1 Mann an Brustkrebs und 1 Frau an Brust- oder Eierstockkrebs.

Sie müssen nicht immer in algorithmische Systeme oder digitale Anwendungen übersetzt werden. Deren Stärke ist die Einfachheit und Nachvollziehbarkeit.

4.3.2 Bilderkennung bei der Mammografie

Seit einigen Jahren werden künstlichen neuronalen Netzen, die mit überwachten Lernverfahren trainiert werden (Kap. 2.3.2), große Potenziale u. a. für die Befundung von Mammografieaufnahmen unterstellt (Becker et al. 2017; Behrends 2018; Bitkom 2015; Dhungel et al. 2017; Kooi et al. 2017; Lotter et al. 2017; McKinney et al. 2020; Ribli et al. 2018; Weiden 2018). Dazu gibt es vielfältige Forschungs- und Entwicklungsaktivitäten, erste Verfahren kommen zwar schon in Anwendungsnähe, in der regulären medizinischen Praxis sind sie jedoch noch nicht ankommen. Um die Potenziale der verfügbaren Datenbestände und die Herausforderungen bei der Integration komplexer algorithmenbasierter Entscheidungs(unterstützungs)systeme realitätsnäher abschätzen zu können, wird zunächst das derzeitige Vorgehen bei der Mammografiebefundung im Rahmen des nationalen Screeningprogramms skizziert. Dort entstehen zum einen seit Jahren zunehmend große mammografische Datenbestände. Zum anderen müssten entsprechende algorithmische Systeme zur Entscheidungsunterstützung in die Abläufe dieses hochgradig qualitätsgesicherten Programms integriert werden.

Vorgehen im nationalen Mammografiescreeningprogramm

2005 wurde auf Beschluss des Bundestages das nationale Mammografiescreeningprogramm als ein zusätzliches Element der Brustkrebsfrüherkennung eingeführt. Es ist ein Angebot an alle Frauen im Alter von 50 bis 69 Jahren. Sie können im Zweijahresrhythmus eine prophylaktische Mammografieaufnahme ihrer Brust anfertigen lassen, um mögliche Gewebeveränderungen früh erkennen und ggf. behandeln zu können (die Krankenkassen tragen die Kosten [§ 25 SGB V]). Das Programm basiert auf den Empfehlungen der europäischen Leitlinie für die Qualitätssicherung des Mammografiescreenings (EK 2003) und wird in Deutschland fachlich konkretisiert durch

- die Richtlinie des Gemeinsamen Bundesausschusses über die Früherkennung von Krebserkrankungen (Krebsfrüherkennungs-Richtlinie/KFE-RL),
- die interdisziplinäre S3-Leitlinie für die Früherkennung, Diagnostik, Therapie und Nachsorge des Mammakarzinoms (DKG et al. 2021) sowie
- den Bundesmantelvertrag – Ärzte, Anlage 9.2: Versorgung im Rahmen des Programms zur Früherkennung von Brustkrebs durch Mammografie-Screening (BMV-Ä Anl. 9.2).

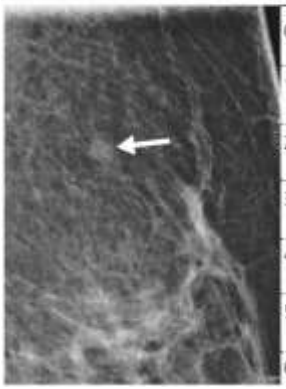
Screeningeinheiten sind auf Brustkrebsfrüherkennung und Spezialdiagnostik spezialisierte, von den Kassenärztlichen Vereinigungen akkreditierte, ambulante Facharztpraxen. Sie werden von einer Radiologin bzw. einem Radiologen (Programmverantwortliche/r) geführt und haben ein Einzugsgebiet von ca. 0,8 bis 1 Mio. Einwohnern, sodass sichergestellt wird, dass dort tätige Radiolog/innen jährlich mindestens 5.000 Mammografien befunden. Die Screeningeinheiten nutzen zertifizierte Praxisinformationssysteme mit standardisierten Patientenakten, die mit den Bildarchiven der Aufnahmegeräte verknüpft sind.

Im Standardverfahren des Screeningprogramms ist kein direkter Kontakt zwischen Ärzt/innen und Programmteilnehmer/innen vorgesehen (er kann auf Wunsch angemeldet werden). Notwendige Informationen erhalten Frauen mit der Einladung über ein Merkblatt oder online.[105] Darin wird darüber informiert, dass keine direktidentifizierenden Daten weitergegeben werden, anonymisierte medizinische Daten jedoch regelmäßig zentral ausgewertet werden, um die Qualität des Programms zu überwachen. Einwilligungen für darüber hinaus gehende Datenweiterverwendungen werden nicht eingeholt. Die Aufnahmen werden von zertifizierten Röntgenassistent/innen gemacht und im Anschluss durch zwei zertifizierte Radiolog/innen unabhängig voneinander visuell befundet (örtlich und räumlich getrennt, keine Kenntnis des anderen Befunds). Diese müssen zum einen erkennen, ob die Qualität der Aufnahme für eine Befundung ausreicht (andernfalls muss eine weitere Aufnahme ggf. mit einem anderen Verfahren gemacht werden). Zum anderen müssen sie unauffällige Aufnahmen (normale Gewebestrukturen und sicher gutartige Gewebeauffälligkeiten [Läsionen]) von auffälligen Aufnahmen (Gewebe mit unklaren oder möglicherweise bösartigen Läsionen) unterscheiden (Abb. 4.2).

Die/der Programmverantwortliche führt die Ergebnisse der fachärztlichen Doppelbefundung zusammen. Wurden die Aufnahmen von beiden als unauffällig befundet, gibt es keine zusätzliche Prüfung, das Ergebnis wird der Frau innerhalb von 7 Tagen schriftlich mitgeteilt. Wurde die Aufnahme mit zusätzlichem Abklärungsbedarf befundet, wird der Fall in einer wöchentlich durchzuführenden Konsensuskonferenz kollegial mindestens zu Dritt beraten. Die/der Programmverantwortliche befundet abschließend. Wird der Befund für auffällig gehalten, wird die betroffene erneut eingeladen. Im direkten Arzt-Patientinnen-Gespräch werden der Befund (lediglich ein Tumorverdacht, keine abgesicherte Diagnose) erläutert und die nächsten Schritte zur Abklärungsdiagnostik vereinbart. Dazu gehören weitere Aufnahmen mit bildgebenden Verfahren, ggf. wird mit einer minimalinvasiven Biopsie eine Gewebeprobe entnommen, die in spezialisierten Pathologielaboren zunehmend automatisiert aufbereitet und fachärztlich mit Mikroskop visuell befundet wird (betroffene Gewebeart, Ausbreitungsmuster, Invasivität des Tumors, Kernteilungsrate usw.). Erst dadurch wird die finale Tumordiagnose gestellt.

105 www.mammo-programm.de (20.11.2021)

Abb. 4.2 Radiologische Befundung von Mammografien

Mammografieaufnahme	Klassifikation der Befunde mittels Breast Imaging Reporting and Data System (BI-RADS®)		
	Nr. Beurteilung/Klassifikation	Befund	Anmerkungen
	0 keine Aussage möglich		weitere Aufnahme nötig (z. B. Vergleichsmammografie, MRI, Sonografie)
	1 normal, negativ, unauffällig (Malignomwahrscheinlichkeit 0%)	unauffällig	keine weitere Abklärung erforderlich, Kontrolle im Routineintervall
	2 sicher gutartige Läsion (z. B. einfache Zyste)	unauffällig	keine weitere Abklärung erforderlich, Kontrolle im Routineintervall
	3 Läsion unklar, wahrscheinlich gutartig (Malignomwahrscheinlichkeit < 2%)	auffällig, Abklärungsbedarf	kurzfristige Kontrolle (6 Monate), eventuell Biopsie
	4 Läsion unklar, eher malign (Malignomwahrscheinlichkeit 2–95%)	auffällig, Abklärungsbedarf	Biopsie/Gewebeprobe
	5 Läsion sicher malign (Malignomwahrscheinlichkeit > 95%)	auffällig, Abklärungsbedarf	Biopsie/Gewebeprobe
	6 Malignität durch Biopsie nachgewiesen		Dokumentation (Register) und Therapie

Quelle: Heindel et al. 2021, S. 132; Sickles et al. 2013

2018 haben 2,9 Mio. Frauen am Screeningprogramm teilgenommen. 4 % wurden wiedereinbestellt, bei 1 % wurde eine Biopsie genommen, bei der Hälfte (16.300 Fälle) wurde ein Tumor diagnostiziert und eine operative Entfernung empfohlen (Kääb-Sanyal/Hand 2020, S. 6).

Maßnahmen zur Qualitätssicherung

Diverse Maßnahmen sollen die Qualität sowohl der einzelnen Arbeitsschritte auf der Ebene der Screeningeinheit (intern) als auch des gesamten Programms auf regionaler und auf nationaler Ebene absichern (Überblick z. B. in Vomweg 2017). In den Anhängen zum Bundesmantelvertrag werden diese Maßnahmen konkretisiert sowohl hinsichtlich der Datenerfassung (u. a. zur Prüfung der Bildqualität, zu Formaten der Datenspeicherung, zur Prozessierung von Bilddatensätzen sowie zum Einsatz von Bildverarbeitungsalgorithmen) als auch zur Befundung (Anhang 5 BMV-Ä Anl. 9.2). Am Screeningprogramm teilnehmende Radiolog/innen müssen ihre fachliche Befundungsbefähigung jährlich nachweisen. Dafür müssen sie 50 Mammografien aus der zentralen Fallsammlung[106]

106 Screeningeinheiten schicken dem zuständigen regionalen Referenzzentrum regelmäßig geeignete pseudonymisierte Mammografien, die diese an die zentrale Kooperationsgemeinschaft des Screeningprogramms weitergeleiten. Letztere führt alle Mammografien zu einer zentralen Fallsammlung zusammen. Diese Fallsammlung soll Mammografien mit ausreichender Bildqualität von mindestens 500 Frauen enthalten, wobei das Spektrum mammografisch relevanter Erkrankungen zu berücksichtigen ist, alle Aufnahmen auffälliger Befunde durch Gewebeproben histopathologisch abgesichert sein müssen und ein Teil der Mammografien in der Fallsammlung regelmäßig ausgetauscht wird (Anhang 5 BMV-Ä Anl. 9.2). Die Kooperationsgemeinschaft stellt den Referenzzentren anonymisierte Fallsammlungen zur Prüfung der Befundbefähigung zur Verfügung.

befunden und dabei eine Sensitivität (auffällige Befunde werden richtig erkannt) und Spezifität (unauffällige Befunde werden richtig erkannt) von mindestens 90 % erreichen.

Zudem wird die Qualität des Mammografiescreenings anhand definierter Kennziffern (u. a. Anteil positiver Befunde der bildgebenden und der histopathologischen Untersuchungen, Anteile falsch positiver und falsch negativer Befunde) vereinfacht dargestellt und darüber mess- und vergleichbar gemacht (BMV-Ä Anl. 9.2). Derartige kennzahlengetriebenen Analysen werden auch als *Benchmark* bezeichnet (Kap. 4.4.2). Definierte Kennziffern werden quartalsweise für jede/n Radiolog/in und für jede Screeningeinheit mit einen im PIS integrierten Analysetool automatisiert ermittelt, visualisiert und anschließend einrichtungsintern diskutiert (Teil des internen Controllings). Diese QS-Datensätze haben Bezüge zu Fachärzt/innen und Screeningeinheiten, in der Regel nicht zu Patient/innen[107] und werden nur programmintern verwendet. Für das programmweite Controlling übermittelt jede Screeningeinheit einen definierten arztpseudonymisierten, einrichtungsbezogenen QS-Datensatz dem zuständigen regionalen Referenzzentrum, das alle QS-Datensätze zusammenführt und sie um definierte Leistungsabrechnungs- und Registerdaten der jeweiligen Kassenärztlichen Vereinigungen ergänzt (Kap. 5.3). Die Ergebnisse dieser externen Qualitätssicherung werden mit jeder Screeningeinheit einzeln beraten. Die Regionalzentren übermitteln arzt- und einrichtungsanonymisierte QS-Gesamtdatensätze an die zentrale Kooperationsgemeinschaft Mammografie. Sie sind eine Datenbasis für die jährliche Gesamtevaluation des Screeningprogramms (Kääb-Sanyal/Hand 2020, S. 9).

Nutzenbewertung des Mammografiescreeningprogramms

Ähnlich wie die Qualität wird auch der Nutzen des Programms anhand von Kennziffern dargestellt und bewertet. Als wichtigste Kennziffer zur Nutzenbewertung gilt die langfristige Mortalität (Anzahl Todesfälle einer Personengruppe in einem bestimmten Zeitraum). Sie soll perspektivisch anhand der epidemiologischen Krebsregister (Kap. 4.1.4) für die Gruppe der programmteilnehmenden Frauen und für die der nichtteilnehmenden ermittelt und verglichen werden. Aussagekräftige Kennziffern zur Langzeitmortalität können erst ab 2021 ermittelt werden (Heindel et al. 2021; Vomweg 2017). Bis dahin wird das Programm anhand von Surrogatkennziffern bewertet. Dazu gehören die Diagnoserate, das Tumorstadium bei Erstdiagnose, die Falsch-Positiv-Raten sowie Überdiagnosen. In den ersten Jahren des Programms stiegen die Brustkrebsdi-

107 Wenn eine befundende Ärztin im Quartal nur einen falsch positiven Befund erstellte, könnte die betroffene Patientin von Dritten mit Zusatzwissen re-identifiziert werden.

agnosen bzw. die Brustkrebsentdeckungsrate wie erwartet an, im weiteren Verlauf vielen diese Werte wieder. Derzeit erhalten 6 von 1.000 Frauen, die am Programm teilnehmen, eine Tumordiagnose, in 5 von 6 Fällen handelt es sich um einen invasiven Tumor, der zunehmend im Frühstadium entdeckt wird. Bei 1 von 6 Fällen wird durch die Mammografie eine Gewebeveränderung festgestellt, die sich im späteren Verlauf zu einem Tumor entwickeln kann, aber nicht muss. Zur Sicherheit wird zur Behandlung geraten, d. h. operative Entfernung der Gewebeveränderung und ergänzende systemische Behandlungen, um das Risiko eines möglichen Rückfalls zu senken.

Programmbefürwortende interpretieren die gestiegenen Diagnoseraten und die zunehmende Entdeckung von Krebs im Frühstadium als Erfolg des Programms, dessen Ziel es ja ist, Krebsfrühstadien zu entdecken und Heilungschancen zu erhöhen (z. B. Heindel et al. 2021; Kettritz 2018, S. 4 ff.). Kritiker/innen weisen ebenfalls auf die gestiegenen Brustkrebsdiagnosen hin, betonen bei ihrer Bewertung jedoch in besonderem Maße die Falsch-Positiv-Raten und Überdiagnosen, die in Übertherapien münden sowie gesundheitliche Risiken durch unnötige Strahlenbelastung vieler befundfreier Teilnehmerinnen (stellvertretend Kettritz 2018, S. 7 ff.). Da es unterschiedliche Einschätzungen gibt, was als Überdiagnose und als Übertherapie zu bezeichnen ist, gibt es auch unterschiedliche Berechnungen und Bewertungen. Eine diesbezügliche Vereinheitlichung der Definitionen wird gefordert, ist allerdings schwer zu realisieren (Kettritz 2018; Kooperationsgemeinschaft Mammographie 2017).

Die seit Jahren durchgeführten Begleituntersuchungen der nationalen Screeningprogramme können bisher weder die Pro- noch die Kontrapositionen vor allem bezüglich der Überdiagnosen und -therapien eindeutig belegen oder entkräften. Durch die immer längere Zeiträume abdeckenden Nachbeobachtungen wird eine Reduktion der Brustkrebsmortalität bei Programmteilnehmenden von ca. 20 % als zunehmend gesichert angesehen (DKG et al. 2021, S. 49 f.). Grundsätzlich ist das Mammografiescreeningprogramm in Deutschland lediglich ein Angebot zur Früherkennung. Frauen werden ermuntert, Vor- und Nachteile individuell abzuwägen (ggf. in Absprache mit einer Ärzt/in) und eine Programmteilnahme selbst zu entscheiden (G-BA 2017, S. 6 ff.). Insgesamt nehmen gut 50 % der in Frage kommenden Frauen am Screeningprogramm teil (Programmbefürwortende hatten ursprünglich mit einer 70 %en Teilnahme gerechnet).

Computerassistierte Mammografiebefundung: Stand der Technik, Bewertung, Technikdiffusion und Folgedimensionen

Die Vision, die fachärztliche Mammografiebefundung durch algorithmische Systeme zu unterstützen oder sogar teilweise zu ersetzen, gibt es seit langem. Erste Ansätze wurden bereits vor 50 Jahren beschrieben (Winsberg et al. 1967).

Mit der Umstellung auf digitale Aufnahmetechniken erhielt die Vision der automatisierten Objekterkennung neue Impulse. Entsprechende Verfahren werden allgemein auch als CAD-Systeme (computer-assisted detection oder computer-aided diagnosis) bezeichnet (Cheng et al. 2016, S. 244). Einige CAD-Systeme segmentieren lediglich Bildbereiche anhand der Strukturen, vergleichen sie mit Inhalten spezifischer Referenzdatenbanken, markieren und beschreiben auffällige Bereiche (Form, Größe, Struktur). Andere CAD-Systeme gehen einen Schritt weiter und befunden die auffälligen Bereiche mittels fachspezifischer Klassifikationen (z. B. BI-RADS-Klassifikation für die Mammografiebefundung; Abb. 4.2).

1998 hat die US-amerikanische Food and Drug Administration (FDA) die erste auf symbolischen Verfahren aufbauende CAD-Software zur Mammografiebefundung zertifiziert, nachdem deren Sicherheit und Leistung anhand von Testdatensätzen belegt wurde (Sensitivitäts- und Spezifitätsanforderungen, die für spezialisierte Fachkräfte gelten, wurden erreicht) (Lehman et al. 2015). Zudem hat die FDA Zertifizierungsleitlinien für CAD-Verfahren und spezielle Testdatensätze (z. B. public INbreast Dataset) herausgegeben (FDA 2012), um die Transparenz des Zertifizierungsprozesses zu verbessern und die Planungssicherheit für softwareentwicklende Unternehmen zu erhöhen. In den USA sind inzwischen mehrere CAD-Verfahren zur Unterstützung der Mammografiebefundung zertifiziert.

Beim US-amerikanischen Screeningprogramm wird die Erstbefundung in der Regel nur von einer Radiologin bzw. einem Radiologen durchgeführt. Seit 2002 übernehmen die Träger der Screeningprogramme die Kosten für eine zusätzliche CAD-Assistenz, die seitdem zunehmend eingesetzt wird. 2008 wurden in den USA 74 % der Mammografieaufnahmen des Screeningprogramms softwareunterstützt befundet (Lehman et al. 2015). In Großbritannien und in den Niederlanden wurde der ergänzende CAD-Einsatz zusätzlich zur radiologischen Doppelbefundung getestet.

Erste Nutzenbewertungen unter Anwendungsbedingungen zeigten in allen drei Ländern, dass Radiolog/innen mit CAD-Assistenz nur annähernd ähnlich viele Tumore richtig erkannten (maximal gleich gute Sensitivität), insgesamt nicht schneller befundeten, aber häufiger unauffälliges Gewebe als tumorverdächtig bewerteten (schlechtere Spezifität) – also mehr Fälle überdiagnostizierten als Radiolog/innen, die diese Verfahren nicht einsetzten (Gilbert et al. 2008; Lehman et al. 2015; Taylor/Potts 2008; Timmers et al. 2012). Da Mammografiescreeningprogramme ohnehin wegen der Überdiagnosen in der Kritik standen und stehen, verwundert es nicht, dass in Deutschland beim Einsatz dieser CAD-Systeme kein Zusatznutzen im Vergleich zum etablierten fachärztlichen Beurteilungsverfahren anerkannt und sie nicht im Rahmen des nationalen Screeningprogramms eingesetzt wurden (DKG et al. 2012, S. 37).

Diese erste Generation der CAD-Verfahren setzte noch komplexe symbolische Verfahren zur Objekterkennung ein. Inzwischen werden vor allem im Bereich der Bilderkennung künstliche neuronale Netze anwendungsreif. Die Ausgangslage für den Einsatz derartiger KNN scheint beim Mammografiescreening in einigen Punkten besonders gut (Pisano 2020): Durch die nationalen Programme werden Mammografieaufnahmen in großen Mengen digital erzeugt, strukturiert gespeichert (hoher Interoperabilitätsgrad) und langfristig aufbewahrt. Sie sind fachlich befundet, wobei Auffälligkeiten einer überschaubaren Anzahl von Befundklassen zugeordnet werden. Alle auf Mammografieaufnahmen gefundenen Auffälligkeiten werden durch Anschlussuntersuchungen mehrmals geprüft und spezifiziert. Auch werden im Rahmen der Qualitätssicherung bereits Fallsammlungen angefertigt und kontinuierlich aktualisiert. Mit ihnen wird derzeit die Befundbefähigung von Radiolog/innen regelmäßig geprüft. Diese Datenbestände könnten auch zu Trainings- und Prüfungsdatensätzen aufbereitet werden. Ein großer Vorteil von KNN gegenüber den auf symbolischen Verfahren basierten CAD-Systemen der ersten Generation sei, dass sie am Ergebnis selbst trainiert werden und keine Formeln und Modelle mehr benötigen, die das Wissen bzw. die Vorgehensweise zur Tumorerkennung explizit repräsentieren. Die bisher vielversprechendsten Studienergebnisse publizierten McKinney et al. (2020). Sie konnten Aufnahmen von 25.000 Frauen aus dem britischen und von 3.000 Frauen aus dem US-amerikanischen Mammografiescreeningprogramm retrospektiv nutzen, um ein KNN mit einem überwachten Lernverfahren zu trainieren, Auffälligkeiten auf Mammografieaufnahmen zu erkennen und zu markieren sowie zu klassifizieren. Im Rahmen der Studie wurden dann Sensitivität und Spezifität der KNN-basierten CAD-Software mit der der Screeningprogramme und mit der von einzelnen Radiolog/innen verglichen. Im Vergleich zum US-amerikanischen Screeningprogramm mit der dort verankerten Einzelbefundung war das KNN-basierte CAD-Verfahren vergleichbar sensitiv, aber erstmals spezifischer. Im Vergleich zum britischen Screeningprogramm mit Doppelbefundung und Konsensusentscheidung war das KNN-basierte Verfahren statistisch nicht unterlegen (McKinney et al. 2020, S. 93).

McKinney et al. bewerten die Studienergebnisse zurückhaltend und weisen auf vielfältige Unsicherheiten und offene Fragen hin: Bevor das KNN-basierte CAD-Verfahren bei der frühen Brustkrebsdetektion assistierend eingesetzt werden könne, seien weitere Prüfungen unter Alltagsbedingungen nötig. Wie man aus den Erfahrungen mit der ersten Generation von CAD-Verfahren wisse, sind Alltagsbedingungen regelmäßig komplexer und vielschichtiger und in Folge kann die Leistung unter Alltagsbedingungen schlechter sein, als die unter Studienbedingungen ermittelte. Unsicherheiten gibt es auch bezüglich der eingesetzten Trainingsdaten. Unklar ist, inwiefern in der Studienpopulation alle Bevölkerungsteile adäquat berücksichtigt waren oder ob möglicherweise einige

unterrepräsentiert waren, die das KNN in Folge schlechter befunden könne (nichtintendierte Diskriminierung). Weitere Unsicherheiten gibt es, ob sowohl bei 2-D- als auch bei den noch nicht so lange verfügbaren 3-D-Aufnahmeverfahren vergleichbare Leistungen erbracht werden können. McKinney et al. sprechen sich auch dafür aus, mögliche Folgen in Bezug auf die klinischen Arbeitsabläufe innerhalb der nationalen Screeningprogramme in den Blick zu nehmen: Verändern sich die Aufgabenschwerpunkte von Radiolog/innen? Müssen sie Ergebnissen des CAD-Systems mehr Aufmerksamkeit widmen und diese intensiver prüfen? Wie ist zu verfahren, wenn das spezifischere CAD-Programm anders befundet als Fachärzt/innen? Wer trägt die Verantwortung, wer haftet? Welche Folgen würden sich ergeben, wenn die durchgängige Doppelbefundung reduziert werden könnte? Würden dadurch lediglich Effizienzgewinne erzielt, Fachkräftemangel reduziert oder auch Arbeitsprozesse umorganisiert, sodass Ärzt/innen sich intensiver ihren schwerkranken Patient/innen widmen könnten?

In der nationalen Brustkrebsleitlinie schlagen sich die vielfach unterstellten Potenziale von CAD-Verfahren bisher nicht nieder. Sie werden zwar seit Jahren erwähnt (DKG et al. 2012, S. 37 ff.), jedoch wird bisher stets betont, dass CAD-Systeme die Doppelbefundung nicht ersetzen können (DKG et al. 2021, S. 46). Bei der Darstellung technologischer Weiterentwicklungen werden vielmehr die Fortschritte bei der Bildgebung von 2-D- auf 3-D-Mammografie thematisiert. Die Fachgesellschaften bewerten die Kombination von 2-D- und 3-D-Aufnahmeverfahren aufgrund der signifikanten Erhöhung der Detektionsrate (deutlicher Sensitivitätsgewinn und sehr gute Spezifität) als den derzeit vielversprechendsten technologischen Entwicklungsansatz für das Mammografiescreening (DKG et al. 2021, S. 51 f.). Wird die Bildgebung in diese Richtung weiterentwickelt, müssen einerseits völlig neue Trainingsdatensätze bereitgestellt und andererseits CAD-Systeme mit diesen Daten trainiert, getestet und zertifiziert werden. Erste CAD-Entwicklerteams berücksichtigen diese Weiterentwicklung bereits (McKinney et al. 2020; Ribli et al. 2018). Wenn die visuelle Doppelbefundung plus Konsensusentscheidung dadurch noch sensitiver wird, steigen die Leistungsanforderungen an CAD-Systeme.

Im Ausblick weisen McKinney et al. darauf hin, dass diese ersten CAD-Erfolge bei der Mammografiebefundung in etlichen anderen medizinischen Bereichen wahrscheinlich schwerer zu erreichen sind, weil die Datenbasis zum Trainieren von KNN schlechter sei. Auch etliche andere Entwicklerteams sehen die Erstellung umfangreicher qualitativ hochwertiger Trainings- und Testdatensätze als vordringliche Aufgabe und große Herausforderung an (z. B. Ribli et al. 2018; Veta et al. 2014): Denn durch die kontinuierliche Weiterentwicklung sowohl der Aufnahmetechnik (z. B. immer höhere Bildauflösung, Umstellung von 2-D-auf 3-D-Aufnahmeverfahren) als auch des Labelings auffälliger Bildele-

mente (textuelle und klassifizierende medizinische Notationen) müssen Bildersammlungen regelmäßig aktualisiert werden, um sie als Trainings- und Testdatensätze nutzen zu können.

Einschätzung

Die Mammografiebefundung wird seit einigen Jahren als ein möglicher medizinischer Einsatzbereich für Assistenzsysteme diskutiert, die künstliche neuronale Netze einsetzen. Es gibt dazu vielfältige Forschungs- und Entwicklungsaktivitäten. Erste Ansätze rücken in Anwendungsnähe. Der Durchbruch in die Anwendung könnte am ehesten in Ländern mit Screeningprogrammen gelingen, die eine vergleichsweise geringe Befundqualität erreichen. Die Integration in das in Deutschland etablierte Mammografiescreeningprogramm, das als eines der qualitativ besten in Europa gilt (Heindel et al. 2021, S. 134 f.), dürfte schwerer sein (bei bereits realisierter hoher Befundqualität ist ein Zusatznutzen schwerer zu erreichen).

Trainings- und Testdatensätze sind von herausragender Bedeutung für die Entwicklung derartiger Assistenzsysteme. Durch technische Weiterentwicklungen bei den Aufzeichnungsgeräten müssen immer wieder neue Trainingsdatensätze erstellt werden (z. B. Umstellung von 2-D- auf 3-D-Aufnahmetechniken). Die im Rahmen des nationalen Mammografiescreeningprogramms generierten Datenbestände haben aufgrund der hohen Qualität und des erreichten Interoperabilitätsgrades erhebliches Potenzial. Sowohl die Möglichkeiten und Grenzen der Weiterverwendung dieser Daten als auch die schrittweise Integration derartiger Assistenzsysteme in die bestehenden medizinischen Strukturen und Arbeitsabläufe des Mammografiescreeningprogramms sollten eingehender untersucht werden.

4.3.3 Interpretation genetischer Daten für die Therapieplanung

Seit vielen Jahren zielen umfangreiche Forschungsanstrengungen darauf ab, die ablaufenden biologischen Prozesse bei bösartigen Gewebeneubildungen besser zu verstehen. Diesbezüglich müssen bei der Analyse genetischer Daten zwei Besonderheiten beachtet werden: Zum einen gibt es kein Durchschnittsgenom. Jedes Gen kommt in vielfachen Varianten vor und jedes Gen ist in unterschiedlichen Zellen in unterschiedlichem Maße aktiv. Genetische Veränderungen in somatischen Zellen können individuell nur über den Vergleich mit gesunden Zellen der gleichen Person festgestellt werden. Zum anderen können mit der sich ständig weiterentwickelnden Sequenzier- und Datenspeichertechnologie genetische Daten in immer größerer Detailgenauigkeit aus einzelnen Zellen ausgelesen werden. Die entstehenden Datensätze sind sehr groß. Um die geneti-

schen Veränderungen bei der Entstehung und Entwicklung von Krebserkrankungen zu verstehen, werden seit Jahren weltweit immer genauere genetische Daten bis auf molekularer Ebene aus einzelnen gesunden und veränderten Zellen von Krebspatient/innen ausgelesen, verglichen, Abweichungen erfasst und kartiert und so schrittweise tumorspezifischer *genetischer Fingerabdrücke* erstellt. Spezifische datenanalytische Verfahren sind zwingend erforderlich, um Auffälligkeiten, Übereinstimmungen oder Muster zu finden und wiederzuerkennen. Erkenntnisse aus der Grundlagenforschung finden zunehmend ihren Weg in die Anwendung, u. a. zur spezifischeren Therapieplanung oder genaueren Krankheitsprognostik.

Prognostische Multigentests

Bei Brustkrebserkrankungen werden seit Jahren im Rahmen klinischer Studien zusätzlich zur klassischen Befundung und Diagnostik umfangreiche genetische Daten erfasst, die Behandlung dokumentiert und der Krankheitsverlauf über längere Zeit beobachtet. Ein Data-Mining-Ansatz beruht auf der Idee, anhand des Krankheitsverlaufs die Aggressivität des Tumors zu klassifizieren und im Anschluss in den genetischen Datensätzen der Patientinnen der einzelnen Klassen nach spezifischen Mustern, Häufungen und Unterschieden zu suchen. Mit einem solchen Vorgehen haben u. a. van ’t Veer et al. (2002) 70 genetische Veränderungen und Paik et al. (2004) 21 genetische Veränderungen aus Forschungsdaten in unterschiedlichen Gruppen von Brustkrebspatientinnen herausgefiltert, die sie als einen zusätzlichen Indikator für die Tumoraggressivität bewerten. Diese genetischen Auffälligkeiten sind das primäre Ergebnis des Data-Mining-Prozesses (ähnlich wie ermittelten Alltagskriterien im ersten Anwendungsbeispiel, die ein erhöhtes Erkrankungsrisiko indizieren). Um dieses Data-Mining-Ergebnis im Rahmen der Behandlung nutzbar machen und auf neue Patientinnen anwenden zu können, reicht keine Checkliste. Vielmehr hat jedes Team einen *Multigentest* entwickelt, der anzeigt, ob die jeweils definierten brustkrebstypischen genetischen Veränderungen in den Tumorzellen neuer Patientinnen auftreten und wenn ja, wie aktiv sie sind. Anhand des Testergebnisses könne auf die voraussichtliche Entwicklung des Tumors der neuen Patientin geschlossen und das Rückfallrisiko nach erfolgreicher Brustkrebsoperation abgeschätzt werden. Das Testergebnis könne als ein zusätzlicher Indikator zur Prognose der Krankheitsentwicklung herangezogenen werden und die Entscheidung für oder gegen eine ergänzende Chemotherapie nach erfolgreicher vollständiger operativer Entfernung eines Mammakarzinoms im Frühstadium unterstützen (Abb. 4.3).

Abb. 4.3 Schematische Darstellung prognostischer Multigentests

Eigene Darstellung

Kern dieser Multigentests sind Microarrays mit unterschiedlichen Feldern. Jedes Feld ist mit einem spezifischen Genfragment bestückt. Für den Test werden aus dem Tumorgewebe einer Patientin erst einzelne Zellen und dann die darin enthaltene mRNA isoliert. Letztere wird mit einem Floureszenzfarbstoff markiert. Die markierte mRNA wird mit dem Array zusammengebracht. Komplementäre mRNA bindet an einzelne Felder des Arrays, der Rest wird ausgewaschen. Eine hochauflösende Laserkamera nimmt die Intensität und die Wellenlänge der Farbe jeder Position auf – es entsteht ein Genexpressionsmuster. Das erzeugte Muster einer neuen Tumorzelle wird mit den jeweiligen Originalmustern verglichen und anhand der Übereinstimmungen bzw. Abweichungen auf die Existenz und Aktivität definierter genetischer Veränderungen und in Folge auf das Rückfallrisiko der Patientin geschlossen (Abb. 4.3).

Derartige Tests sind rechtlich In-vitro-Diagnostika bzw. Medizinprodukte hoher Risikoklassen (Kap. 4.2). Für die Zertifizierung reichte es bisher, die Leistung anhand retrospektiver Datenanalysen nachzuweisen. Dazu wird von Patientinnen, die vor Jahren an Brustkrebs erkrankten und bei denen man die Behandlungsform und die Rezidiventwicklung kennt, rückwirkend das Genexpressionsmuster anhand des in Biobanken konservierten Tumormaterials analysiert und das Rückfallrisiko bewertet und im Anschluss geprüft, inwiefern diese Bewertung aus heutiger Sicht richtig war. Zudem starten prospektive klinische Studien, in denen in einer Gruppe der Test zusätzlich zur etablierten Risikoabschätzung anhand klinischer Prognosefaktoren (Alter der Patientin, Tumorgröße Befunde der Gewebeuntersuchungen) eingesetzt wurde und in der anderen nicht. Anhand der weiteren Krankheitsentwicklung können dann Leistung und Nutzen dieserart datenbasierter Tests prospektiv bewertet werden. Solche Studien dauern jedoch meist sehr lange (Rezidive entstehen oftmals erst nach Jahren).

Zwar sind unterschiedliche Multigentests zur Bewertung der Tumoraggressivität sowohl in den USA als auch in Europa und Deutschland zertifiziert und marktverfügbar, der Nutzen dieserart Risikoklassifizierung wird jedoch seit Jahren kontrovers zwischen Testanbietenden, medizinischen Fachgesellschaften, Prüfeinrichtungen und Kostenträgern diskutiert. Dadurch war und ist die

Integration dieser auf komplexen Annahmen und Datenanalysen beruhenden Tests in medizinische Versorgung ein langer Weg. In den ersten Jahren kam es sowohl auf die behandelnde Einrichtung an, ob sie einen Multigentest für angemessen hielt und einen Kostenübernahmeantrag stellte, als auch auf die jeweilige Krankenkasse, ob sie die diesem Antrag stattgab (ausführlich in Wilkens 2017). Auch die Beauftragung des Instituts für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) brachte keine Klarheit bezüglich des Zusatznutzens etlicher zugelassener Tests. Vielmehr ist ein jahrelanger Methodenstreit um die Qualität einzelner Studien und die Aussagekraft von Zwischenergebnissen entbrannt. Anders als die US-amerikanischen und britischen Institutionen, die vorgelegte Studien und der Ergebnisse weitgehend akzeptierten und einen Zusatznutzen anerkannten, bewertete das IQWiG die Qualität der meisten vorgelegten klinischen Studien als unzureichend und sah keine ausreichend verlässlichen Belege für die Anerkennung eines Zusatznutzens (IQWiG 2016, 2018, 2020). Auch nationale Fachgesellschaften haben sich intensiv mit den methodischen Vorgehensweisen und den Argumenten zu den jeweiligen Bewertungen auseinandergesetzt. Bisher lassen sie eine Empfehlung offen (DKG et al. 2021, S. 126 ff.). Der G-BA als oberstes Entscheidungsgremium hat 2020 eine Kostenübernahme in engen Grenzen beschlossen (G-BA 2020). In der Begründung wird darauf hingewiesen, dass die Tests nur Wahrscheinlichkeiten über ein Rückfallrisiko ermitteln und nur ergänzende Informationen für Fachärzt/innen liefern, eine Behandlungsempfehlung jedoch in deren Verantwortungsbereich liegen. Die Testergebnisse könnten weder eindeutig klären, welche Frau eine Chemotherapie benötigt noch liefern sie verlässliche Vorhersagen, ob eine bestimmte Frau tatsächlich einen Rückfall haben wird oder nicht noch nicht.

Einschätzung

Die Suche nach Strukturen und Mustern in genetischen Daten ist analytisch anspruchsvoll und komplex. Besondere Herausforderungen ergeben sich, wenn derartige Muster für prognostische Verfahren verwendet werden, die relativ weit in der Zukunft liegende Ereignisse antizipieren. Betrachtet man den Data-Mining-Prozess im Anwendungsbeispiel im engeren Sinn, führte er lediglich zu Informationen über genetische Veränderungen, die bei aggressiven Brustkrebsformen häufig – aber nicht immer – auftreten. Diese Ergebnisse wurden zunächst fachlich diskutiert und geprüft (Data-Mining zur Wissenserweiterung).

Gesellschaftliche Herausforderungen und Folgen ergeben sich erst, wenn jenseits der Wissenserweiterung auch mögliche Anwendungsszenarien in den Blick genommen werden (Nutzung von Data-Mining-Ergebnissen). Im Anwendungsbeispiel mündeten die Data-Mining-Ergebnisse in die Entwicklung von Medizinprodukten bzw. in-vitro-Diagnostika hoher Risikoklassen. Für diese

müssen gesundheitsbezogene Sicherheit, Leistungsfähigkeit und Nutzen im Rahmen der Produktentwicklung, also unter Laborbedingungen bzw. mittels klinischer Studien, nachgewiesen werden (3 Hürden für die Zertifizierung bzw. das Inverkehrbringen). Für die Aufnahme in die GKV-Leistungskataloge werden zudem (Zusatz-)Nutzenbelege unter Anwendungsbedingungen gefordert (4. Hürde des ersten Gesundheitsmarktes). Wie bereits bei CAD-Systemen zur Bildbefundung deutlich wurde, sind dafür national eigenständige Standards und Qualitätsanforderungen maßgeblich. Die in Deutschland über Jahre geführte kontroverse Debatte zur Nutzenbewertung von prognostischen Multigentests hat gezeigt, dass es national unterschiedliche Positionen und Meinungen gibt, welche methodischen Standards gelten und erfüllt werden sollten, um derartige Medizinprodukte hoher Risikoklassen in die Erstattungsfähigkeit der GKV zu bringen. Dazu müssen bisher explizite prospektive klinische Studien durchgeführt werden. Wenn es gelänge, entsprechende Daten aus den Krebsregistern zeitnah bereitzustellen, könnten diese eine zusätzliche Datengrundlage zur Nutzenbewertung darstellen.

Um Herstellern von datenbasierten Werkzeugen zur Diagnose und Behandlung von Krankheiten mehr Planungssicherheit im Innovationsprozess zu geben, werden nicht nur für die Zertifizierung, sondern auch für die Aufnahme in den GKV-Leistungskatalog Leitlinien, einheitliche Qualitätsanforderungen und methodische Standards gefordert. Im 2020 verabschiedeten Digitale-Versorgungs-Gesetz wurden für Medizinprodukte niedriger Risikoklassen bereits Verfahren definiert und Prozesse gestrafft, sodass diese schneller in die Regelanwendung gelangen können. Für risikoreiche datenbasierte Werkzeuge zur Diagnose und Behandlung von Krankheiten sollten die Prozesse zur Integration in die Regelversorgung ebenfalls gestrafft werden.

Neben der Kontroverse um den Zusatznutzen(nachweis) einzelner Tests wird mitunter auch kontrovers diskutiert, dass Verfahren in den GKV-Leistungskatalog aufgenommen werden, in deren Rahmen genetische Daten von GKV-Versicherten bei Firmen außerhalb der EU erhoben, gespeichert und (weiter)verwendet werden (Beerheide 2019).

Tumor-Genomsequenzierung – Big Data in Forschung und Anwendung

Im Rahmen des 2008 gestarteten International Cancer Genome Consortium (ICGC) arbeiten führende öffentlich finanzierte Krebs- und Genomforschungseinrichtungen aus 16 Ländern, darunter das Deutsche Krebsforschungszentrum (DKFZ) und das NCT gemeinsam daran, die somatischen Anomalien auf genetischer, epigenetischer und transkriptomischer Ebene (alle DNA- und RNA-Veränderungen) von 50 Krebsarten (darunter Brustkrebs) jeweils anhand von 500 Tumorpatient/innen nach einheitlichen Standards zu kartieren. Diverse

Richtlinien definieren die notwendigen Qualitätskriterien, sichern hohe Datenschutzstandards, die schnelle organisatorische Datenfreigabe und den freien Zugang über Treuhandstrukturen für qualifizierte Forscher/innen sowie die freie Verfügbarkeit von publizierten Analyseergebnissen über die ICGC-Internetseiten ab (Open Data und Open Access in der medizinischen Forschung). Die Erwartungen an dieses Projekt waren und sind groß: Wichtige onkogenetische Veränderungen und deren Entwicklungsstufen sollen aufgedeckt, klinisch relevante Subtypen für Prognosen und individualisiertere Therapieformen definiert sowie bessere Wege zur Diagnose, Behandlung und Prävention vieler Krebsarten entwickelt werden.

Die Datenerhebung des ICGC läuft seit mehr als 10 Jahren. Die ursprünglich favorisierte IT-Architektur mit weltweit 8 Datenrepositorien, darunter eines in Heidelberg, und die Notwendigkeit des Datendownloads für die Analyse begrenzen jedoch die Nutzung der Genomdatensätze, zumal jedes der durchgeführten Krebskartierungsprojekte mehrere Petabyte große Datensätze liefert.[108] Nur wenige Institutionen weltweit haben die Kapazität, diese immensen Datensätze für eigene Analysen herunterzuladen. Um den Zugang zur genetischen Datenbank des ICGC zu verbessern, wurden 2015 erstmals ca. 5 % der 25.000 Krebsgenomdatensätze vollständig verschlüsselt in die Cloud von Amazon Web Services gestellt. Für autorisierte Forschende sind direkte Datenanalysen ohne Datendownload über die Seven-Bridges-Plattform möglich, über die Spezialsoftware für die Analyse großer genetischer Datensätze bereitgestellt wird.[109] Inzwischen kann auch eine in Kanada initiierte akademische Forschungscloud für die Speicherung und Analyse eines weiteren Teildatensatzes genutzt werden.

Anhand dieser Daten können unterschiedliche Wissenschaftskonsortien mit speziellen mathematischen Verfahren (u. a. Mutationssignaturanalysen) die somatische Genetik unterschiedlicher Krebsformen einschließlich epigenetischer und transkriptomischer Prozesse zunehmend besser nachvollziehen. Für Brustkrebs wurden durch den Vergleich von gesunden und Tumorzellen von mehr als 500 Patientinnen bisher 3,5 Mio. substituierte Basenpaare und 78.000 genetische Umstrukturierungen herausdestilliert (Brinkman et al. 2019). Das Konsortium erstellte ein Register aller beobachteten somatischen Mutationen von 21 Brustkrebsarten. 900 Veränderungen bewerten sie als möglicherweise relevant für die Tumorentwicklung, 93 seien hochrelevant (sogenannte Treibermutationen). Bei unklarer Befundlage (gutartige Zyste oder bösartige Gewebeveränderung) könnte die Existenz solcher Treibermutationen die Brustkrebsdiagnose auslösen. Die entstehenden ICGC-Tumormutationsregister mit den jeweiligen Bewertungen sind ein Kernelement für neue Formen der datenbasierten

108 1 Petabyte entspricht dem Datenvolumen von 223.000 DVDs.
109 www.sevenbridges.com (10.11.2021)

molekularen Tumordiagnostik. Sie bilden den Datengrundstock für spezialisierte Krebsdiagnostikdienstleister:

Molecular Health GmbH (2004 in Heidelberg gegründet) nutzt die Tumormutationsregister und erweitert die Datenbank kontinuierlich um einschlägige Studienergebnisse zur Wirksamkeit und zu Nebenwirkungen von diversen therapeutischen Ansätzen, die sie von Fachkräften bewerten und zuordnen lassen (Karlberg 2018). Die von der Firma entwickelte cloudbasierte Plattformtechnologie »Molecular Health Guide« können Fachärzt/innen bzw. Tumorboards nutzen, um den genetischen Datensatz eines neuen Tumorpatienten mit den Registerdaten zu vergleichen, interaktiv zu analysieren und sich einen Report mit Therapieoptionen erstellen zu lassen. Dieser Report enthält laut Hersteller u. a. evidenzbasierte Informationen zu Biomarkern, Arzneimitteln und ihren Wechselwirkungen sowie Hinweise zu passenden klinischen Studien. Laut Herstellerangaben würde technisch im Grunde nichts anderes gemacht werden, als das, was Ärzt/innen tagtäglich tun – nur systematischer und schneller. 2017 wurde dieser Ansatz mit dem deutschen Leader Award »Sonderpreis Disruption« ausgezeichnet.[110] Das Verfahren zur automatisierten Erkennung und Klassifikation von genetischen Mutationen ist als In-vitro-Diagnostikum zertifiziert. Die digital unterstützte Interpretation der Analyseergebnisse und die Reports mit Therapieoptionen für behandelnde Ärzt/innen gehen über die eigentliche Befundung genetischer Daten hinaus und können als medizinisches Entscheidungsunterstützungssystem aufgefasst werden. Inwiefern dieses medizinisch nützlich ist, wird untersucht. Im März 2018 startete der Anbieter zusammen mit der Charité und zwei GKK ein erstes Projekt zur Nutzenbewertung in Bezug auf Therapieempfehlungen für Kinder, die nach der ersten Tumorstandardtherapie ein Rezidiv erlitten. Die genetische Diagnostik wird an der Charité durchgeführt, die GKK tragen die Kosten. Die Analyseergebnisse werden in einer ersten Sitzung des Tumorboards bewertet und das weitere therapeutische Vorgehen besprochen und vereinbart. Danach werden mit dem »Molecular Health Guide« die Daten aus der Tumorgenomsequenzierung analysiert und der Report mit möglichen Therapieoptionen maschinell erstellt. In einer zweiten Sitzung des Tumorboards werden die Ergebnisse und Therapieempfehlungen aus beiden Vorgehensweisen miteinander verglichen. Inzwischen hat sich die Nützlichkeit dieses Verfahrens sowohl bei pädiatrischen Tumoren als auch bei Eierstockkrebs bestätigt, bei anderen Krebsarten laufen Untersuchungen. Erste individuelle Erstattungsverträge zwischen einzelnen Krankenhäusern und Krankenkassen werden vereinbart. Dies ist der erste Schritt in die GKV-Versorgung.

Sophia Genetics S.A. (2011 in der Schweiz gegründet)[111] nutzt einen ähnlichen Ansatz. Auch diese Firma ergänzt die Daten der Tumormutationsregister

110 www.molecularhealth.com/de/2017/07/01/charite-tk-und-molecular-health/ (10.11.2021)

111 https://www.sophiagenetics.com/company/about-us/ (10.11.2021)

kontinuierlich um Zusatzinformationen (u. a. zu Patogenitätsklassen, Therapieoptionen, med. Publikationen) und nutzt eine firmeneigene Cloud, mit der in den Daten aus der Tumorgenomsequenzierung von Krebspatient/innen nach bekannten genetischen Mutationen gesucht werden kann. Zur genauen genetischen Diagnostik gehört auch die Zuordnung zu vordefinierten Patogenitätsklassen. Die dafür eingesetzten Algorithmen sind als In-vitro-Diagnostika zertifiziert. Bei der Patogenitätszuordnung erreichen sie laut Herstellerangaben eine nahezu 100 %ige Sensitivität und Spezifität. Auch dieser Dienstleister bietet zusätzliche Hinweise zu erfolgversprechenden Therapieoptionen. Zudem können auch bisher unbekannte genetische Mutationen in die Datenbank aufgenommen werden, um sie perspektivisch bei der Interpretation weiterer Datensätze berücksichtigen zu können. Laut Anbieter nutzen den Diagnoseservice inzwischen mehr als 1.000 Krankenhäuser in 77 Ländern (Stand November 2021). Genetische Datensätze von Hunderttausenden Patient/innen wurden bereits prozessiert. Für die jeweiligen Krankenhäuser sinkt der Aufwand für die Befundung genetischer Datensätze erheblich.

Von den medizinischen Fachgesellschaften wird zunehmend anerkannt, dass sich bei sehr seltenen Krebserkrankungen oder wenn bewährte Behandlungsmöglichkeiten ausgeschöpft sind, durch Tumorgenomanalysen zusätzliche Therapieansätze ergeben können (dkfz 2020). Als »Neue Untersuchungs- und Behandlungsmethode« sollen sie schrittweise in das Leistungsabrechnungssystem des stationären Bereichs integriert werden.

Einschätzung

Der kontinuierliche Ausbau von Datenbeständen, die Anreicherung durch zusätzliche Informationen, die Klassifikation einzelner Elemente, der Abgleich und die Bewertung neuer Datensätze sind typische Vorgehensweisen digitaler Geschäftsmodelle. Das Marktsegment der Analyse genetischer Daten wird dafür zunehmend relevant. Wie in etlichen Bereichen der Plattformökonomie, wird derjenige, der die umfangreichsten Datenbestände aufbauen, klassifizieren und mit Zusatzinformationen anreichern kann, die besten Möglichkeiten haben, mittels automatisierter Analysen zusätzliche Informationsdienste für behandelnde Ärzt/innen bzw. medizinische Einrichtungen entwickeln und anbieten zu können. Diese Services sind mehr als bloße In-vitro-Diagnostika. Man bezeichnet sie teilweise als medizinische Entscheidungsunterstützungs- oder Assistenzsysteme. Erste Systeme zur Tumorspezialdiagnostik schaffen den Schritt von der experimentellen Anwendung in die Versorgung.

4.3.4 Medizinische Assistenzsysteme

Die Vision der Entwicklung von Computerprogrammen, die unter Verwendung von Patientendaten und medizinischen Wissens Ärzt/innen bei ihrer Arbeit unterstützen, gibt es seit vielen Jahren (Shortliffe 1987). Ursprünglich wurden sie als medizinische Expertensysteme bezeichnet und grob in drei Kategorien unterteilt (Gamper/Steimann 1996):

- *Informationsmanagementsysteme* zielen einerseits auf eine benutzerfreundliche Verwaltung und Aufbereitung von Daten und Informationen (PIS/KIS; medizinische Bibliografien), sodass diese von medizinischen Fachkräften schnell erfasst werden können. Andererseits sollen damit administrative Prozesse unterstützt werden.
- *Systeme zur Fokussierung der Aufmerksamkeit* sollen auffällige Werte erkennen und hervorheben (z. B. Laborsysteme), bei der Medikation und beim Behandlungsablauf assistieren (z. B. auf Kontraindikationen hinweisen) oder durch Überwachungs- und Erinnerungsfunktionen die Therapietreue [Compliance] unterstützen.
- *Entscheidungsunterstützungssysteme* zielen drauf ab, anhand patientenbezogener Daten und einer medizinischen Wissensbasis nicht nur Analyseergebnisse zu bewerten, sondern auch Diagnosen vorzuschlagen, beste Behandlungsoptionen zu berechnen und/oder geeignete Therapien zu empfehlen.

Die ersten beiden Kategorien gelten seit Jahren als technisch machbar, teilweise als realisiert, auch wenn sie kontinuierlich weiterentwickelt werden. Im Routineeinsatz sind seit Jahren z. B. weitgehend automatisiert ablaufende Programme zur EKG- und Blutgasanalyse oder zum Test von Lungenfunktionen (Pfeffer 2002).

Entscheidungsunterstützungssysteme zielen darauf ab, nicht nur bei Einzelaktivitäten wie z. B. bei der Bildbefundung zu unterstützen, sondern auch komplexere medizinische Aufgaben zu realisieren. Sie gleichen Daten eines neuen Patienten mit krankheitsspezifischen Daten- und Wissensbeständen ab. Dafür nutzen sie unterschiedliche analytische Verfahren. Die Spanne reicht von relationalen Verknüpfungen (klassische symbolische Verfahren) über Wahrscheinlichkeitsnetze bis zu maschinellen Lernverfahren und trainierten KNN. Sie gelten als präziser als Universalsuchmaschinen wie z. B. Google (Tab. 4.1).

Tab. 4.1 Programme zur Unterstützung med. Entscheidungen (Auswahl)

Name	Assistenz bei	Anmerkungen	Leistungs-bewertung
MYCIN (1970er Jahre) Nachfolger: *GIDEON* (kein Routine-einsatz)	Therapie: Antibi-otikaauswahl bei Infektionen	erste Software dieser Art; nutzte maschinelle Lern-verfahren	70% Tref-ferquote
INTERNIST (1979) Nachfolger: *CADUCEUS*	Diagnose: innere Medizin (1.000 Krankhei-ten)	nutzte ähnliches masch-inelles Lernverfahren wie MYCIN	
DXplain (1980er Jahre)	Diagnose: Allgemeinmedi-zin (2.400 Krankheiten)	relationale Verknüpfung von 5.000 Symptomen/Be-funden mit Krankheiten mit Wahrscheinlichkeits-ranking	3,45* (im Mittel)
Mole-Analyzer (1998 für Ärzte/in-nen)	Diagnose: Haut-krebsvorsorge	Klassische symbolische Verfahren zur Bilderken-nung u. -bewertung	80 bis 88%
Isabel (2000 für Ärzt/innen; 2012 für Patient/innen)	Diagnose: Symptom-checker	semantische Suchmaschine anhand med. Lehrbücher werden zu Symptomen Di-agnosen gesucht	3,45* (im Mittel)
Phenomizer (2009 für Ärzt/innen)	Diagnose: seltene Erkran-kungen	relationale Verknüpfung von Bildern, genetischen Daten und Symptomen (Human Phenotype Onto-logy – HPO) mit 7.500 Erb-krankheiten, priorisiert wahrscheinlichste Diagno-sen	
Face2Gene (2015 für Ärzt/innen)	Diagnose: seltene Erkran-kungen	nutzt masch. Lernverfahren weist Patientenfotos mögli-che HPO-Symptome zu	

Watson for Oncology (2014 für Ärzt/innen)	Diagnose/ Therapie: Onkologie	semantische Suchmaschine sucht in med. Bibliografien und Patientenakten Ähnlichkeiten zu neuen Fällen	
Ada (2016 für Patient/innen)	Diagnose: Symptom-checker Allgemeinmedizin	nutzt maschinelle Lernverfahren, stellt zielgerichtete Anamnesefragen und schlägt nächste Schritte ggf. Arztbesuche vor	besser als andere Apps

* Score: 0 bis 4: falsch bis sehr ähnliche Diagnose, 5: richtige Diagnose

Quellen: Bitkom 2015, S. 69 ff.; Bond et al. 2012; Gäßner 2002; Karlberg 2018; Köhler et al. 2009; Lenzen-Schulte 2017; Lüdtke 2002; Puppe 2014; Ross/Swetlitz 2017; https://ada.com/de; www.fdna.com; www.medaware.com; www.moleanalyzer.com (10.11.2021)

Die erste Generation von Expertensystemen lief auf Großrechnern und basierte auf umfangreichen Abfragen und Dateneingaben (z. B. MYCIN, INTERNIST; Tab. 4.1). Die zweite Generation erleichterte die Interaktion mit grafischen Benutzeroberflächen bereits deutlich (z. B. DXplain). Um den Benutzeraufwand weiter zu senken, setzt die dritte Generation derzeit teilweise auf Spracherkennung und semantische Suchmaschinen[112] (z. B. Isabel, Watson). Einige neuere Services bieten mobile Anwendungen (z. B. Isabel, PhenIX). In der Regel werden Entscheidungsunterstützungssysteme für Ärzt/innen konzipiert. Einige Entwickler/innen sprechen mit ihren Services zum Teil auch medizinische Laien an (Direct-to-Consumer-Services, wie z. B. Ada) quasi als niedrigschwellige Erstfilter, die Symptome checken und bei der Entscheidung assistieren, ob Ärzt/innen konsultiert werden sollten.

Die Watson-Initiative von IBM

2012 startete IBM medienwirksam die Watson-Initiative. Nachdem die semantische Suchmaschine »Watson« in einer US-amerikanischen Quizshow natürlichsprachliche Anfragen besser beantwortet hatte als die menschlichen Quizteilnehmer, wollte IBM u. a. mit »Watson for Oncology«[113] die Krebsbehandlung weltweit revolutionieren. In einer Kooperation mit dem New Yorker

112 Semantische Suchmaschinen versuchen, den Sinn (Semantik) natürlichsprachlicher Anfragen zu erfassen und suchen in einem definierten Datenbestand nach passenden Antworten. Teilweise auch als kognitive Systeme oder kognitives Computing bezeichnet. Behauptet wird, dass der Computer Sprache versteht und z. B. die Relevanz von Fachartikeln für eine Diagnose erkennen und Therapieempfehlungen nebst Begründung liefern könne (Bitkom 2015).

113 www.mskcc.org/blog/mskcc-and-ibm-will-collaborate-powerful-new-medical-technology (10.11.2021)

Memorial Sloan Kettering Cancer Center (MSKCC), eine der renommiertesten US-amerikanischen Krebskliniken, begann man deren Wissensbestände, festgehalten in Patientenakten, med. Unterlagen, klinikeigenen Behandlungsleitlinien und Kontraindikationen bis hin zu Dokumentationen und Datensätzen aus zahlreichen klinischen Studien und fachspezifischen Publikationen aufzubereiten und einzulesen. Um sich bei einem neuen Tumorfall Rat bezüglich Diagnose und Behandlung holen zu können, muss dessen Patientenakte in vorgegebenen Formaten eingelesen werden. Zur genauen Methodik und den eingesetzten Verfahren der Datenverarbeitung, die dann Diagnose- und Behandlungsempfehlungen generieren, ist öffentlich nicht viel bekannt. Laut Bitkom (2015, S. 69 ff.) wird die semantische Suchmaschine darauf trainiert, mündlich geäußerte Fragen von Ärzt/innen sinngemäß zu erkennen und dann Empfehlungen gegebenenfalls mit Unsicherheiten und Begründungen auszugeben. IBM bietet die Nutzung des Assistenzsystems Krankenhäusern weltweit an (Nutzungsgebühr pro Tumorfall je nach Leistungsumfang zwischen 200 und 1.000 US-Dollar) (Ross/Swetlitz 2017). Teilnehmende Kliniken übermitteln Patientenakten mit möglichst umfangreicher Krankengeschichte, Gesundheitszustandsbeschreibung und gegebenenfalls existierenden Behandlungspräferenzen. Je nach vereinbartem Leistungsumfang erhalten sie Diagnose- und Therapieempfehlungen mit möglicherweise existierenden Begründungen und Fachpublikationen. Im Standardfall basieren die generierten Empfehlungen auf den von der New Yorker Klinik eingelesenen Daten, Dokumenten und deren Behandlungsleitlinien.

Wenn in den vergangenen Jahren datenanalytische Verfahren mit einem Potenzial für große Veränderungen in der Medizin thematisiert wurden, wurde regelmäßig auch die IBM-Watson-Initiative genannt (z. B. Akademien der Wissenschaften Schweiz 2015, S. 28 f.; Bitkom 2015, S. 69 ff.). Denn für zwei allgemeine Herausforderungen im medizinischen Alltag wurde eine technische Lösung versprochen: Eine adressiert das allgemeine Big-Data-Problem, dass immer größere Datenbestände in unterschiedlichen Formaten generiert werden. Die andere adressiert die angeblich rasante Zunahme medizinischen Wissens (regelmäßig veranschaulicht anhand der Flut medizinischer Publikationen durch PubMed-Neueinträge). Protagonist/innen argumentieren, dass beide Datenbestände maschinell schneller gescannt und verarbeitet werden können, als durch Menschen. Der Nutzen für eine bessere Behandlung von gravierenden Krankheiten läge dann auf der Hand. Eine maschinelle Analyse spare Ärzt/innen Zeit und ermögliche auch in Krankenhäusern mit begrenztem eigenem Fachkräftestab eine schnelle und medizinisch hochwertige Diagnose und Behandlungsplanung. Kritiker/innen wiesen darauf hin, dass für »Watson for Oncology« weder Leistungsnachweise noch der postulierte Nutzen belegt und bewertet wurde. Aufgrund der fehlenden Nutzenbelege befragten Ross/Swetlitz (2017) weltweit Fachkräfte, die im IBM-Watson-Team arbeiten, die maschinelle Lernverfahren entwickeln und die Tumorpatient/innen behandeln und

»Watson for Oncology« im klinischen Alltag ausprobierten, nach ihren Einschätzungen. Diese fielen wohl auch aufgrund der anfangs geschürten hohen Erwartungen eher ernüchternd aus. Die ursprünglich anvisierten großen Veränderungen in der Krebsbehandlung konnten bisher noch nicht realisiert werden (He 2020; Jie et al. 2021). Mitglieder des IBM-Watson-Teams gaben zu, dass das Assistenzsystem noch immer in den Kinderschuhen stecke. Der Aufwand, den medizinische, datenverarbeitende und IT-Expert/innen seit Jahren erbringen, um »Watson for Oncology« in einzelnen Krebsarten kontinuierlich auf dem neuesten Stand des Wissens zu halten, sei nach wie vor erheblich.

Kritiker/innen vermuten inzwischen, dass die Daten weit weniger intelligent verarbeitet werden, als die IBM-Marketingabteilung suggeriert. Zwar werden maschinelle Lernverfahren zur Spracherkennung eingesetzt, vermutet wird aber, dass die Diagnose- und Therapieempfehlungen weniger auf maschinellen Lernverfahren und kontinuierlichem Training beruhen, sondern vielmehr die vorgegebenen medizinischen Algorithmen der Behandlungsleitlinien des New Yorker Krankenhauses wiedergeben. Mitunter wird »Watson for Oncology« als »MSKCC in einer tragbaren Box« bezeichnet mit dem Problem, dass die Situation einer US-amerikanischen Hightech-Klinik nicht unbedingt mit der von Kliniken in anderen Ländern übereinstimmt. Insbesondere Kliniken in anderen Industrieländern stellen den Nutzen dieser Vorgehensweise in Frage. Zum einen gibt es dort vielfältige hochspezialisierte Fachkräfte, die gerade bei häufigen Krebsarten über umfangreiches Wissen verfügen und die tendenziell eher bei unklaren Befundlagen und seltenen Tumorerkrankungen zusätzliche Assistenz als nützlich bewerten dürften. Vorrangiges Ziel der Watson-Entwicklung sei es aber, ein Assistenzsystem zur Behandlung der häufigsten Krebsarten bereitzustellen. Zum zweiten behandeln diese Kliniken nicht nach US-amerikanischen Leitlinien. Ross/Swetlitz (2017) verweisen auf Pilotprojekte in niederländischen und dänischen Kliniken, bei denen »Watson for Oncology« nur in 33 % der Fälle zu den gleichen Diagnose- und Therapieempfehlungen kam wie die jeweiligen Tumorboards und man in Folge dort auf das Assistenzsystem wieder verzichtete. Aus Südkorea wurde berichtet, dass »Watson for Oncology« Tests und Behandlungen empfahl, die dort teilweise gar nicht zugelassen oder nicht im Leistungskatalog der nationalen Krankenversicherungen enthalten waren. Etwas positiver fielen die Einschätzungen des Klinikpersonals aus Thailand oder der Mongolei aus, die zumindest die schnellen Rückmeldungen und die jeweiligen Begründungen nützlich fanden. Jedoch seien die Therapieempfehlungen auf nordamerikanische Tumorpatient/innen abgestimmt, nicht auf die jeweils heimische Bevölkerung mit teilweise abweichenden Stoffwechselprozessen. Mengenangaben für Therapeutika seien nicht einfach übertragbar. Eine erste diesbezügliche systematische Literaturrecherche in den weltweit relevantesten medizinischen Literaturdatenbanken und die Meta-Analyse der extrahierten 9 Publikationen aus dem asiatischen Raum bestätigte diese Einschätzungen

im Wesentlichen (Jie et al. 2021): Je schwerwiegender die Krebserkrankungen waren, desto weniger stimmten die Behandlungsempfehlungen, die Watson for Oncology gab, mit denen von Tumorboards überein.

In den Industrieländern werden die fehlenden Leistungs- und Nutzenbewertungen bei Assistenzsystemen mit dieser medizinischen Tragweite als höchst kritisch angesehen. Rechtlich sind es zertifizierungspflichtige Medizinprodukte höchster Risikoklasse (Kap. 4.2). Der experimentelle Einsatz nichtzugelassener Medizinprodukte entspricht außerhalb von genehmigungspflichtigen klinischen Studien nicht einer dem anerkannten Stand der medizinischen Wissenschaft entsprechenden Behandlung. Für die Zertifizierung müssten Hersteller Sicherheit, Leistung und Nutzen der Verfahren nachweisen. Für derartig breite Ansätze gibt es bisher kaum methodische Vorgaben.

Assistenzsysteme zur Erkennung seltener Erkrankungen

Bei seltenen Erkrankungen ist die derzeitige Gesamtkonstellation zur Diagnostik und Behandlung ungünstiger als bei weit verbreiteten Krankheiten: Symptome und Befunddaten sind oftmals diffuser, die bestehende Wissensbasis kleiner, Leitlinien nur begrenzt erarbeitet. Da diese Krankheiten selten auftreten, begegnen primärversorgende Ärzt/innen diesen nur sehr selten. Statt eine umfangreiche Spezialdiagnostik anzuordnen, werden Symptome häufiger vorkommenden Krankheiten zugeordnet und zunächst versucht, diese zu therapieren. Oft dauert es vergleichsweise lange, bis wegen ausbleibenden Therapieerfolgs hochspezialisierte Fachärzt/innen eingeschaltet werden und selbst diesen fällt es oft nicht leicht, seltene Erkrankungen sofort richtig zu diagnostizieren (Gillessen-Kaesbach et al. 2016). Ein Assistenzsystem, das auf eine möglicherweise vorliegende seltene Erkrankung hinweist, könnte diesen Zeitraum bis zur Diagnose möglicherweise verkürzen und damit einen gesundheitsbezogenen Nutzen generieren.

Unterschiedliche Teams arbeiten an Algorithmen, die Ärzt/innen frühzeitig auf möglicherweise vorliegende seltene Erkrankungen aufmerksam machen und damit mögliche Fehlbehandlungen zu Beginn vermeiden und den Zeitraum bis zur richtigen Diagnose verkürzen sollen. Das Assistenzsystem »Phenomizer« nutzt die »Human Phenotype Ontology« (HPO), eine an der Charité entwickelte Onlinedatenbank, in der mehr als 10.000 Symptome mit 7.500 seltenen Erbkrankheiten verknüpft sind. Der Algorithmus durchsucht und vergleicht genannte Symptome mit den in der Datenbank enthaltenen Daten und gibt eine Liste mit möglicherweise vorliegenden seltenen Erkrankungen aus. Der Nutzen dieses Vorgehens wird mit der Zeitersparnis assoziiert. Die Liste soll behandelnden Ärzt/innen helfen, die Krankheit schneller einzukreisen, zielgerichteter vorzugehen und ggf. eine genetische Spezialdiagnostik durchzuführen (Köhler et al. 2017).

Face2Gene setzt bereits existierende Verfahren zur Gesichtserkennung ein und weist anhand von Patientenfotos möglicherweise vorliegende HPO-Symptome zu. Um den Trainingsdatensatz aufzubauen, wurden Menschen mit diagnostizierten seltenen Erkrankungen gebeten, Fotos (Gesicht, Füße, Hände) bereitzustellen. Mit maschinellen Lernverfahren werden KNN trainiert, Besonderheiten zu erkennen, die auf möglicherweise vorliegende Erbkrankheiten hinweisen. Nach diesen Besonderheiten wird dann auf Bildern von neuen Patientinnen gezielt gesucht. Ärzt/innen erhalten Hinweise für weitere spezialdiagnostische Untersuchungen.[114]

Die beispielhaft skizzierten Ansätze, die vom Erscheinungsbild einzelner Patient/innen – dem Phänotyp – ausgehend Auffälligkeiten und Symptome zuordnen, werden teilweise auch als »deep phenotyping« bezeichnet. Durch eine Typisierung des klinischen Erscheinungsbildes soll der Kreis von möglicherweise vorliegenden genetischen Veränderungen zielgerichtet eingegrenzt werden. Die Sicherheit und Leistungsfähigkeit dieser unterschiedlichen technologischen Lösungsversuche wird nach wie vor geprüft (Stand November 2021).

Auch im Rahmen der Watson-Initiative hat man begonnen, seltene Erkrankungen als mögliches Einsatzgebiet zu sondieren. Ein umfangreicher englischsprachiger Wissensbestand wurde in den IBM-Watson-Explorer eingelesen. Im Oktober 2016 startete u. a. ein Pilotprojekt in Deutschland (Rhön-Klinikum/ IBM 2016). Am Zentrum für unerkannte und seltene Erkrankungen der Marburger Uniklinik wurde ein Fragebogen entwickelt, den neue Patient/innen digital beantworteten. Die Antworten wurden zusammen mit der im Vorfeld zu digitalisierenden Patientenakte und dem Anamnesebogen ohne direkt personenbezogene Merkmale an das Watson-System in der IBM-Cloud geleitet. Aus den teilweise in natürlicher Sprache auf Deutsch formulierten Antworten und den Daten der Patientenakte und des Anamnesebogens wurden zunächst wesentliche Informationen extrahiert, übersetzt und im »Cognitive Core« mit dem englischen Wissensbestand verglichen und eine Liste von fachlich belegbaren Hypothesen zur Diagnosefindung zurückgesendet. Nach einem hoffnungsvollen Beginn (Rhön-Klinikum 2017), wurde die Zusammenarbeit Ende 2017 beendet. Die Technik war für den Krankenhausalltag schlicht unbrauchbar gewesen (Balzter 2018).

Ada

Das Berliner Start-up Ada-Health[115] begann 2011 mit der Entwicklung eines Assistenzsystems, das anhand von geschilderten Symptomen wahrscheinlichste Diagnosen vorschlägt. Ursprünglich sollte es Ärzt/innen unterstützen. 2016 än-

114 www.face2gene.com (10.11.2021)
115 https://ada.com/de/ (10.11.2021)

derte man den Fokus und konzipierte den Symptomchecker als App für medizinische Laien. Technische Grundlage ist eine medizinische Datenbank mit Symptombewertungen im Millionenbereich, die kontinuierlich ausgebaut wird. Laut Hersteller werden Symptome von neuen Fällen über einen dynamischen Fragenkatalog erfasst, dessen Kern ein KI-Verfahren bildet, das kontinuierlich trainiert wird, relevante weitere Symptome zu checken, wahrscheinliche Diagnosen einzugrenzen, nächste Schritte und ggf. ärztliche Konsultationen vorzuschlagen. Ziel der App sei es, allgemeine Internetrecherchen nicht aber ärztliche Konsultationen zu ersetzen. Vergleichstests mit anderen Apps zur Symptombeurteilung und mit telefonischen Konsultationen bei Hausärzt/innen zeigten, dass Ada zum einen besser war als andere symptomcheckende Apps und zum anderen Symptome ähnlich bewertet wurden wie in telefonischen Hausarztkonsultationen (Gilbert et al. 2020). Die App ist in Europa als Medizinprodukt der Risikoklasse 1 zertifiziert und wird inzwischen in sieben Sprachen angeboten. Ende 2018 startete das Unternehmen mit zwei namhaften Stiftungen eine Global-Health-Initiative, um in Ländern mit niedrigem und mittlerem Einkommen den Zugang zur Gesundheitsversorgung zu verbessern. Dadurch wurde u. a. auch eine App-Version in Suaheli erstellt. Ziel sei es, den Menschen mit begrenztem Zugang zu medizinischen Versorgungsstrukturen Gesundheitsberatungen anzubieten und Gesundheitspersonal vor Ort zu unterstützen (Ada 2018).
In unterschiedlichen Ländern wurde Ada nach dem Start schnell als eine der besten Gesundheits-Apps bewertet. In Deutschland startete u. a. eine gesetzliche Krankenkasse eine Kooperation und integrierte die App. 2019 erhielt sie bei den German Innovation Awards Gold in der Kategorie »Excellence in Business to Consumer«. Die Prüfung der Einhaltung von Datenschutzstandards kann dabei keine herausragende Rolle gespielt haben. Denn zum einen war die Nutzung der App an die Datenübermittlung an US-amerikanische Tracking- und Analysedienstleister geknüpft. Zum anderen war Ada im Vergleich zu diversen anderen Gesundheits-Apps diejenige, die an die meisten Drittfirmen Daten übermittelte. Aufgedeckt wurde dies von einem investigativen Journalisten, nicht von der zuständigen Datenschutzaufsichtsbehörde (ausführlich Tremmel et al. 2019). Der Hersteller hat nach Bekanntwerden dieser datenschutzbezogenen Schwächen reagiert und die Datenübermittlung an Dritte beschränkt. Die GKK hat die Kooperation beendet und statt eines algorithmenbasierten Chatbots ein telemedizinisches Ärztezentrum eingerichtet und eine Kommunikations-App entwickelt, mit der Versicherte via Textchat oder (Video-)Telefonie mit den dort tätigen Ärzt/innen kommunizieren können.[116]

116 www.tk.de/techniker/magazin/themen/spezial/das-magazin-2-19/tk-doc-app-mit-doktortitel-2074916 (10.11.2021)

Gesamteinschätzung Assistenzsysteme

Medizinischen Assistenz- oder Entscheidungsunterstützungssystemen wird teilweise ein erhebliches Potenzial unterstellt, Ärzt/innen nicht nur bei Einzelaktivitäten wie der Bildbefundung, sondern auch bei komplexen Aktivitäten zu unterstützen, bei denen sie viele Symptome und Befunde wissensbasiert zu Krankheitsdiagnosen und bei Therapieentscheidungen zusammenführen. Den Sprung in die breite Anwendung haben sie bisher kaum geschafft. Die Gründe dafür sind vielfältig (Castelvecchi 2016; Gamper/Steimann 1996; Gäßner 2002; Laursen 2016; Puppe 2014): Teilweise wird bezweifelt, dass ärztliches Vorgehen anhand von vorgegebenen Verfahrensschritten sowie Patient/innen anhand von Messwerten adäquat abgebildet werden können (Stichworte: Komplexitätsreduktion, Kochbuchmedizin). Auch würde die Datenerhebung und -aufbereitung immer mehr Ressourcen binden und die Zeit für den direkten Patientenkontakt kontinuierlich beschneiden (Dokumentationsflut). Teilweise gibt es Befürchtungen, zum Handlanger von Software degradiert zu werden (Kompetenzverlust). Teilweise werden allgemeine Argumente gegenüber maschinellen Lernverfahren genannt (intransparente Funktionsweise in Kombination mit fehlenden Indikatoren für die Richtigkeit im Einzelfall; Beeinflussbarkeit durch falsches Training), die zu Entscheidungsdilemmata führen (Behandlungsentscheidungen würden getroffen, ohne dass Begründungen, Risikofaktoren etc. dargestellt werden) und Fragen zur Arzthaftung aufwerfen. In der Summe können die Gründe für die bisherige Ablehnung mit einem nicht ersichtlichen Nutzen bei der täglichen Arbeit und einer überwiegenden Skepsis insbesondere bei erfahrenen Ärzt/innen zusammengefasst werden. Etliche Ressentiments ließen sich wahrscheinlich überwinden, wenn die Ergebnisqualität der Programme überzeugen könnte. In Bereichen, in denen Fachkräfte ohnehin an ihre Grenzen kommen und/oder es gesundheitsbezogene Lücken oder Defizite gibt, dürfte es leichter sein, einen gesundheitsbezogenen (Zusatz-)Nutzen zu generieren.

Datenbasierte Systeme, die sich direkt an Patient/innen wenden und Arztkonsultationen teilweise ersetzen wollen, sind eine neue Servicekategorie (Stichwort Telemedizin).[117]

117 Telemedizinische Anwendungen werden vom TAB in einem eigenständigen Projekt thematisiert: www.tab-beim-bundestag.de/de/untersuchungen/u40600.html (10.11.2021)

4.4 Administrative Daten: Basis von Geschäftsprozessen

Medizinische Versorgungsprozesse werden im öffentlichen Gesundheitssystem im Kern durch eine Dreiecksbeziehung geprägt. Zu ihr gehört neben der Arzt-Patienten-Beziehung im Rahmen der Behandlung auch die Geschäftsbeziehung zwischen medizinischen Einrichtungen und den Trägern des nationalen Sozialversicherungssystems (vor allem die gesetzlichen Krankenkassen – GKK).[118] Vervollständigt wird die Dreiecksbeziehung durch das Versicherungsverhältnis zwischen Patient/innen und ihren Krankenkassen (als Versicherungsträger). Medizinische Einrichtungen rechnen Leistungen, die sie für gesetzlich Versicherte Patient/innen erbringen, direkt mit deren jeweiliger Krankenkasse ab.

Die Rechts- und Geschäftsbeziehungen zwischen medizinischen Einrichtungen und GKK werden durch das SGB V sowie durch diverse Einzelgesetze, nachgelagerte Ausführungsverordnungen, Verträge und Vereinbarungen im Rahmen der Selbstverwaltung definiert. Grundsätzlich muss die medizinische Versorgung der Versicherten ausreichend und zweckmäßig sein. Sie muss in der fachlich gebotenen Qualität und wirtschaftlich erbracht werden und sie darf das Maß des Notwendigen nicht überschreiten (§ 12 SGB V). Diese Vorgaben begrenzen das Spektrum medizinischer Leistungen, die medizinische Einrichtungen zu Lasten der GKV erbringen und gesetzliche Krankenkassen erstatten müssen (erster Gesundheitsmarkt). Die Einhaltung dieser Vorgaben wird auch mit umfangreichen Datenanalysen überwacht. Unterschiedliche Institutionen sind daran beteiligt. In dieser Konstellation agieren medizinische Einrichtungen als Wirtschaftsbetriebe mit aufgabenspezifischen Sonderstrukturen und kollektivvertraglichen Verpflichtungen.

Medizinische Einrichtungen sind zur Erfüllung zahlreicher öffentlicher Aufgaben gesetzlich verpflichtet, definierte Datensätze zu erstellen und an unterschiedliche Institutionen des Gesundheitssystems weiterzuleiten. Für diese

118 Die medizinische Versorgung wird vor allem über Krankenversicherungen finanziert, entweder die gesetzliche Krankenversicherung (GKV; Rechtsgrundlage: Sozialgesetzbuch V – SGB V) oder private (PKV; Rechtsgrundlage: Gesetz über die Beaufsichtigung der Versicherungsunternehmen [Versicherungsaufsichtsgesetz – VAG], Gesetz über den Versicherungsvertrag [Versicherungsvertragsgesetz – VVG]). Auch die gesetzliche Unfallversicherung (GUV; Rechtsgrundlage: SGB VI), gesetzliche Rentenversicherung (GRV; Rechtsgrundlage: SGB VII) und in Einzelfällen andere Sozialversicherungen decken medizinische Leistungen finanziell ab. Jede Versicherung hat eigene Träger, die GKV z. B. mehr als 100 gesetzliche Krankenkassen (GKK). Dadurch sind das nationale Gesundheitssystem und dessen Datenhaltung stark fragmentiert. Auch wenn die unterschiedlichen Träger jeweils eigene Leistungsspektren sowie Abrechnungs- und Governancestrukturen haben, ist der durch die Sozialgesetzbücher definierte Rahmen bezüglich des Umgangs mit Daten ähnlich. Diese Fallstudie konzentriert sich wesentlich auf die datenbezogenen Strukturen und Analysemöglichkeiten, die durch die GKV geprägt werden.

Datenzusammenstellung benötigen sie ihre jeweiligen PIS/KIS, die einen großen administrativen Arbeitsbereich mit vielfältigen Funktionalitäten dafür haben. Trotz dieses Arbeitsbereichs sind viele Einzelaktivitäten arbeitsintensiv. Die Datenflüsse zu unterschiedlichen Institutionen sowie die primären und sekundären Nutzungsmöglichkeiten und -grenzen werden detailliert vorgeschrieben und komplex reguliert.

4.4.1 Daten zur Leistungsabrechnung

Medizinische Einrichtungen rechnen Behandlungsleistungen mittels definierter patientenbezogener (Leistungs-)Abrechnungsdatensätze ab (Kap. 10 SGB V). Viele Angaben entspringen den Patienten-/Fallakten und werden (um)codiert (Kasten 4.3). Sie werden ergänzt um Abrechnungskennziffern.

Kasten 4.3 Codierungen und Klassifikationen zur Leistungsabrechnung

Patientenbezogene Merkmale werden vor allem mittels Krankenversichertennummer (KV-Nr. 290 SGB V)[119] verschlüsselt, Angaben zu Ärzt/innen mittels Arztnummern (BAN/LANR) und die zu medizinischen Einrichtungen mittels Institutionenkennzeichen (IK). Medizinische Sachverhalte werden mittels statistisch-administrativer Klassifikationen codiert. Derzeit vorrangig relevant sind:

- die *deutsche Modifikation der International Statistical Classification of Diseases and Related Health Problems*, derzeit in der Version 10 (ICD-10 GM) für die Codierung von Diagnosen; sie wird vom DIMDI/BfArM[120] herausgegeben, jährlich fortgeschrieben und erweitert (dadurch verändern sich regelmäßig die Strukturen der Klassifikation; sie hat inzwischen knapp 14.000 Codes);
- der *Operationen- und Prozedurenschlüssel* (OPS) ist die deutsche Modifikation der International Classification of Procedures in Medicine (ICPM) für die Codierung von medizinischen Behandlungsleistungen,

119 Um die Datenzusammenführung unterschiedlicher Lebensbereiche und die Erstellung umfassender Persönlichkeitsprofile durch öffentliche z. T. staatliche Einrichtungen zu verhindern, werden in Deutschland keine allgemeingültigen Personenkennzeichen für alle Administrationsbereiche vergeben. In Folge dürfen z. B. Krankenversichertennummer (§ 290 SGB V) zur Abrechnung der medizinischen Heilbehandlung und Rentenversicherungsnummer (§ 147 SGB VI) zur Abrechnung von rehabilitativen Behandlungsleistungen nicht übereinstimmen.

120 Das ehemals eigenständige Deutsche Institut für Medizinische Dokumentation und Information (DIMDI) ist seit 2020 Teil des Bundesinstituts für Arzneimittel und Medizinprodukte (BfArM).

auch sie wird vom DIMDI/BfArM herausgegeben, jährlich fortgeschrieben und erweitert (inzwischen hat sie mehr als 30.000 Codes);
- national eigenständige *Pharmazentralnummern* (PZN)[121] für Arzneimittelabgaben durch öffentliche Apotheken; es handelt sich um eine von der Informationsstelle für Arzneispezialitäten (IFA GmbH) für 2 Jahre vergebene achtstellige fortlaufende eindeutige Nummerierung.

Erbrachte Leistungen werden anhand von Abrechnungskennziffern pauschaliert vergütet. Im Rahmen der GKV im ambulanten Bereich anhand des Katalogs zum *Einheitlichen Bewertungsmaßstab (*EBM*)* und im stationären Bereich mittels *Fallpauschalenkatalog* (Diagnosis Related Groups – DRG).

Für die (Um-)Codierung sind spezifische Kenntnisse erforderlich: zum einen zu den verwendeten Terminologien und Klassifikationen, die regelmäßig überarbeitet werden, und zum anderen auch spezifische betriebswirtschaftliche, weil die Diagnose- und Behandlungscodierungen die Vergütung bestimmen und auch diese Vergütungskataloge regelmäßig weiterentwickelt werden. Für die Dokumentation, Codierung und Prüfung der Leistungsdaten wurden in den vergangenen Jahren die einrichtungsinternen Kapazitäten kontinuierlich ausgebaut (Bundesrechnungshof 2019, S. 29 f.). Im ambulanten Bereich codieren Ärzt/innen ihre Eintragungen oftmals selbst. In stationären Einrichtungen übernehmen diese Tätigkeit zunehmend speziell ausgebildete medizinische Dokumentations- oder explizite Codierfachkräfte, die an administrativen Arbeitsplätzen der KIS einen selektiven Einblick in die Fallakten haben.

Spezielle Zusatzmodule der PIS/KIS können bei der (Um-)Codierung zunehmend unterstützen: Unter anderem können texterkennende Verfahren bei der Codierung von Freitexteinträgen assistieren (z. B. indem beim Diagnoseeintrag Brustkrebs alle infrage kommenden Schlüsselnummern vorgeschlagen werden) oder erste Plausibilitätsprüfungen vorgenommen werden (z. B. indem geprüft wird, ob für bestimmte Leistungen auch die notwendige Diagnose gestellt wurde). Auch kann mit Simulationsrechnungen bei denen z. B. die Reihenfolge und Anzahl von Haupt- und Nebendiagnosen verändert wird, ermittelt werden, wie sich dies auf die EBM- oder DRG-Zuordnung auswirkt. Im Anschluss können die Abrechnungsdatensätze entsprechend optimiert werden (Bundesrechnungshof 2019, S. 59). Eine gänzlich automatisierte Codierung abrechnungsrelevanter medizinischer Merkmale ist bisher weder möglich noch gewollt. Nach der finalen Festlegung der Diagnose- und Leistungscodes werden im ambulanten Bereich EBM-Ziffern manuell zugeordnet. Im stationären Bereich wird ein Pauschalbetrag pro Behandlungsfall anhand unterschiedlicher

121 Die vom DIMDI herausgegebene deutsche Version der pharmazeutischen ATC-Wirkstoffklassifikation (anatomisch therapeutisch chemisch) mit definierten Tagesdosen (Defined Daily Doses – DDD) wird bisher nicht für die Leistungsabrechnung eingesetzt.

Kennziffern (Haupt- und Nebendiagnosen, Behandlungsleistungen, aber auch demografische Angaben wie Alter, Geschlecht, Postleitzahl) einem zertifizierten Algorithmus automatisiert ermittelt (Kap. 5.2).

Die einzelfallbezogenen Datensätze zur Leistungsabrechnung werden wie folgt weitergeleitet (Schepers et al. 2015, S. 152):

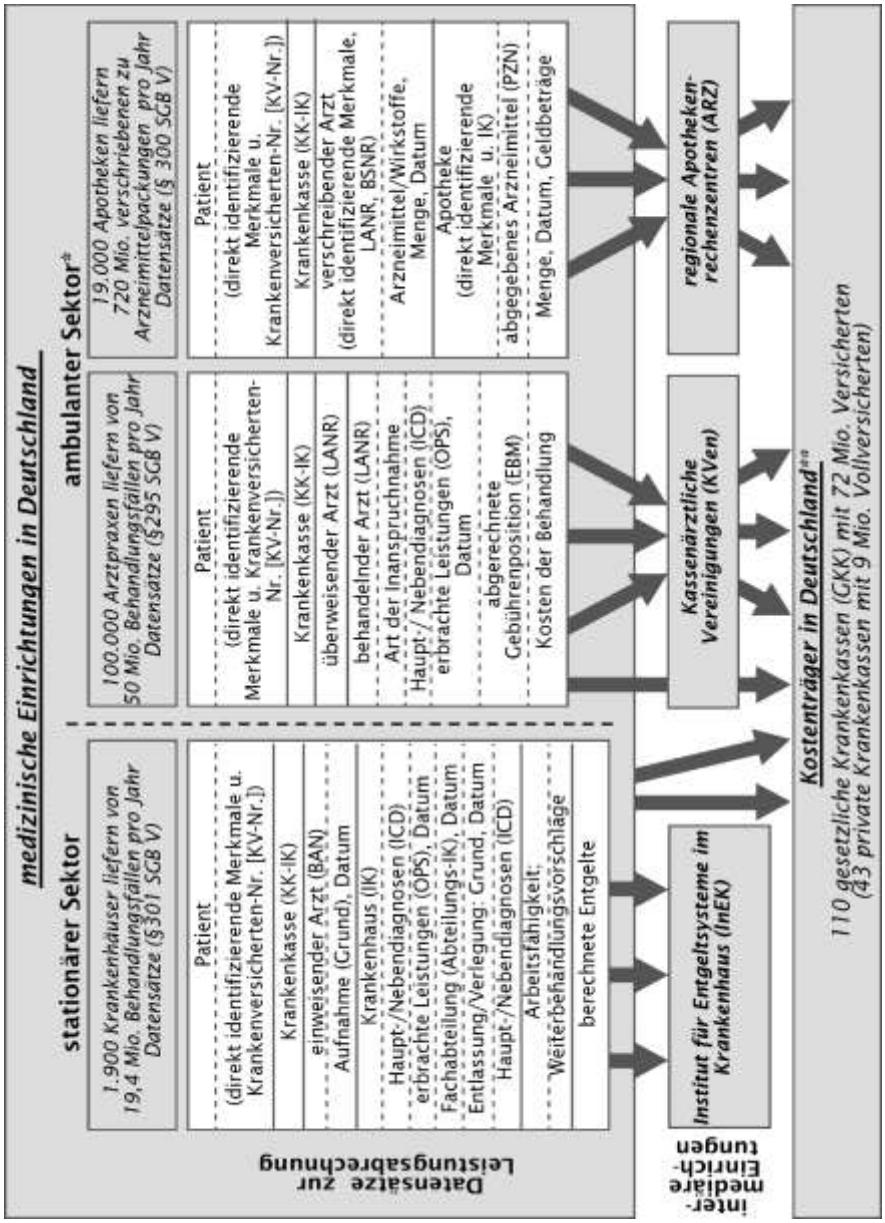
- Krankenhäuser übermitteln nach Behandlungsende einen Datensatz pro Patient/in direkt an die jeweilige Krankenkasse. Außerdem stellen sie einmal pro Jahr alle einzelfallbezogenen Leistungsdatensätze zusammen (§ 21 KHEntgG) und übermitteln diesen Jahresdatenbestand an das Institut für das Entgeltsystem im Krankenhaus (InEK) zur Fortschreibung des Fallpauschalenkatalogs und des Grouper-Algorithmus (Kap. 5.2).
- Arztpraxen stellen quartalsweise einen Datensatz pro behandelte Person zusammen und übermitteln alle Datensätze ihre zuständige Kassenärztliche Vereinigung.[122] Letztere prüfen die Leistungsdaten, ermitteln und realisieren die quartalsweise Vergütung der Praxen, teilen den Datenbestand anschließend kassenweise auf und übermitteln jeder GKK die Leistungsdatensätze ihrer Mitglieder (Kap. 5.3).
- Apotheken übermitteln alle 2 bis 4 Wochen sämtliche eingelöste und vervollständigte Rezepte an privatwirtschaftlich betriebene Apothekenrechenzentren, die die Abrechnung gegenüber den jeweiligen GKK zentral organisieren, den Datenbestand ebenfalls kassenweise aufteilen und jeder GKK die Rezeptdaten ihrer Mitglieder übermitteln (Kap. 5.4).

Abbildung 4.4 stellt die einzelfallbezogenen Leistungsdatenflüsse im Rahmen der GKV-Selbstverwaltung im Überblick dar.

Medizinische Einrichtungen und die dort tätigen Fachkräfte müssen die behandelten gesetzlich Versicherten weder in die Erstellung der Datensätze zur Leistungsabrechnung einbeziehen (nur privat Versicherte erhalten im ambulanten Bereich eine Rechnung) noch über die Datenübermittlung informieren. Gesetzlich Versicherte können der Datenübermittlung nicht widersprechen. Das gesetzlich definierte Verfahren durchbricht die ärztliche Schweigepflicht und beschränkt sowohl die informationelle Selbstbestimmung gesetzlich Versicherter als auch das Rt6ttttttttttt45t4t45tg5t5t445tgg5ttgtg5t6echt auf Wahrung von Betriebsgeheimnissen. Datenempfangende Stellen sind jedoch zur Geheimhaltung verpflichtet und dürfen diese Daten nur für ebenfalls gesetzlich definierte Zwecke verwenden (ausführlicher in Kap. 5).

122 Ärzte, die sich an speziellen Versorgungsformen beteiligen – z. B. hausarztzentrierte Versorgung, integrierte (sektorübergreifende) Versorgung –, rechnen diese Leistungen direkt mit den jeweiligen GKK ihrer Patienten ab. KVen sind an dieser Abrechnung nicht beteiligt.

Abb. 4.4 Leistungsabrechnungsdaten der gesetzlichen Krankenversicherung



BAN: bundeseinheitliche Arztnummer
BSNR: Betriebsstättennummer
EBM: einheitlicher Bewertungsmaßstab
ICD: International Classification of Diseases
zzIK: Institutionenkennzei-chen
LANR: lebenslange Arztnummer
OPS: Operationen- und Prozeduren-schlüssel
PZN: Pharmazentralnummer;
V-Nr.: Versichertennummer

* verkürzte Darstellung; wegen abweichenden Statistiken nicht enthalten:
- Heil-/Hilfsmittelerbringern (Abrechnungsdatensätze nach § 302 SGBV)
- Hebammen (§ 301a SGB V) (rechnen direkt mit Krankenkassen ab)

**verkürzte Darstellung, weitere Sozialleistungsträger fehlen (u. a. Rentenversicherungsträger, Unfallkassen)

Eigene Darstellung auf der Basis des SGB V; Daten: www.destatis.de (10.11.2021)

Alle Datensätze zur Leistungsabrechnung haben Bezüge zu behandelten und behandelnden Personen, zu med. Einrichtungen, Sozialversicherungsträgern sowie zu Zeit und Raum (in der amtlichen Statistik werden derartige einzelfallbezogene Datensätze auch als Mikrodaten bezeichnet). Sie bilden die gesundheitliche Situation der gesetzlich Versicherten, das Vorgehen von Ärzt/innen sowie Geschäftsprozesse medizinischer Einrichtungen ab. Anders als die Behandlungsdaten der Patientenakten sind die zu erstellenden Leistungsdatensätze hochgradig standardisiert. Jenseits der Leistungsabrechnung ergeben sich vielfältige sekundäre Analysemöglichkeiten, insbesondere wenn Datensätze zur Leistungsabrechnung einrichtungsübergreifend verarbeitet werden dürfen. Sowohl für die Leistungsabrechnung als auch für weitere Datenanalysen ist Richtigkeit der Daten von großer Relevanz (Stichwort Data-Dredging; Kap. 2.1).

Richtigkeit von Leistungsabrechnungsdaten

Inwiefern Leistungsabrechnungsdaten die gesundheitliche Situation von Versicherten und den Behandlungsverlauf korrekt abbilden, wird immer wieder kontrovers diskutiert (z. B. Baas/Scherff 2016; Balling 2018; Borchardt 2012; Bundesrechnungshof 2019; Dirschedl 2012). Die Menge der angegebenen (Neben-)Diagnosen und Behandlungsprozeduren steigt seit Jahren. Die Möglichkeiten, diese Daten auf ihre Richtigkeit zu prüfen, sind systemisch begrenzt. Diagnose- und Behandlungsabläufe finden im geschützten Arzt-Patientenverhältnis statt, auch die Dokumentation unterliegt der ärztlichen Schweigepflicht. Gesetzlich Versicherte sind weder an der Behandlungsdokumentation noch am Leistungsabrechnungsverfahren beteiligt. Eine Kontrollfunktion in Bezug auf die Richtigkeit der Daten können Patient/innen kaum wahrnehmen.

Im ambulanten Bereich prüfen Kassenärztliche Vereinigungen teilweise in Kooperation mit Krankenkassen die quartalsweise eingehenden einzelfallbezogenen Leistungsabrechnungsdaten auf Rechtmäßigkeit und Plausibilität (Kap. 5.3). Die aus deren Sicht unkorrekten Positionen werden in der Regel ohne Rücksprache mittels weitgehend automatisierter Verfahren gestrichen, danach die jeweiligen Honorare berechnet. Eine Revision erfolgt nur, wenn Ärzt/innen bei gestrichenen Positionen Widerspruch einlegen.

Im stationären Bereich prüfen die Krankenkassen die Leistungsabrechnungsdatensätze ihrer Versicherten in unterschiedlichem Maße selbst. Der Prüfungsfokus liegt auf kostenrelevanten Angaben, die finanzielle Nachteile für die eigenen Geschäftsprozesse bringen. Nicht alle Krankenkassen prüfen in gleichem Umfang. Zudem haben die jeweils eingesetzten Algorithmen unterschiedliche Treffergenauigkeiten (Bundesrechnungshof 2019, S. 6). Bei Auffälligkeiten werden zum einen die jeweiligen Krankenhäuser um Aufklärung und ggf. um Datenberichtigung gebeten (Vorverfahren). Zum anderen kann der Medizi-

nische Dienst der Krankenversicherung, (MD) mit der Prüfung beauftragt werden (§ 275 SGB V). Nur deren Ärzt/innen dürfen dafür die jeweiligen arztgeführten Patientenakten einsehen (GKK-Angestellte haben keine Einsichtsrechte). Der Aufwand für die Prüfung der Leistungsabrechnungsdaten steigt seit Jahren. Laut Bundesrechnungshof (2019, S. 6) hat der MD 2016 mehr als 2 Mio. Prüfverfahren durchgeführt, also ca. 10 % aller Leistungsabrechnungsdatensätze geprüft (2008 waren es 1,2 Mio.). Bei etwa der Hälfte wurden Auffälligkeiten oder Fehler bei der Codierung und/oder der Abrechnung festgestellt, die, teilweise gerichtlich geklärt werden.

Seit Jahren werden die etablierten Vergütungsverfahren anhand definierter Leistungsabrechnungsdatensätze in Verbindung mit den Prüfungsmöglichkeiten kontrovers diskutiert. Das Leistungsabrechnungssystem wird kontinuierlich stärker ausdifferenziert, dadurch komplexer, aufwendiger und interpretierbarer. Deren Anwendung lässt sich immer schwerer nachvollziehen. Aus finanziellen Erwägungen müssen medizinische Einrichtungen ihre Leistungsabrechnung über die Dokumentation und Codierung im Rahmen des Möglichen optimieren und Kostenträger dies begrenzen, sofern sie dadurch wirtschaftliche Nachteile haben (Bundesrechnungshof 2019, S. 8). Kritiker unterstellen dem Verfahren eine generelle Neigung vor allem zur Überdiagnostizierung, teilweise auch zur Überbehandlung (Baas/Scherff 2016; Dirschedl 2012). Denn medizinischen Einrichtungen wird die Betreuung Schwererkrankter von den GKK tendenziell besser vergütet und gleichzeitig erhalten GKK für schwerkranke Versicherte über den morbiditätsorientierten Risikostrukturausgleich höhere Zuweisungen aus dem Gesundheitsfonds (Kap. 5.5.1). Diese Prozesse binden einerseits immer mehr Ressourcen, andererseits lassen sich systematische Datenverzerrungen nicht ausschließen. Das Ausmaß einer abrechnungsbedingten systematischen Verzerrung von Leistungsabrechnungsdaten, lässt sich kaum sicher abschätzen. In Folge kann die Eignung von Leistungsabrechnungsdaten für bestimmte Data-Mining-Prozesse in Frage und damit die Validität ermittelter Ergebnisse in Abrede gestellt werden (Anwendungsbeispiel Pharmakovigilanz; Kap. 5.5.3).

4.4.2 Daten zur Qualitätssicherung der medizinischen Versorgung

Sowohl Ärzt/innen als auch medizinische Einrichtungen sind zu Maßnahmen verpflichtet, mit denen die Qualität der medizinischen Versorgung gesichert werden soll (§ 630a BGB; §§ 135 bis 139c SGB V; § 2 MBO-Ä). Sie müssen zum einen diverse einrichtungsinterne Maßnahmen zur Sicherung der Strukturqualität durchführen und dokumentieren (Qualitätsmanagement; Anwendungsbeispiel Mammografiescreening; Kap. 4.3.2) und zum anderen an einrichtungsübergreifenden, externen Verfahren zur Qualitätssicherung (QS) teilnehmen.

Umfangreiche Datenanalysen, vor allem Benchmarkanalysen, gehören zunehmend zum Methodenspektrum.

Bei Benchmarkanalysen zur Qualitätssicherung werden unterschiedliche Behandlungsprozesse in medizinischen Einrichtungen zum Untersuchungsobjekt. Dafür wird das nicht direkt messbare Konstrukt »Qualität medizinischer Versorgung« mehrstufig zerlegt und in unterschiedlichen Dimensionen und Teilbereichen anhand vielfältiger Indikatoren vereinfacht dargestellt und mittels Kennziffern quantitativ erfasst. Diese Kennziffern sind meist statistische Indexzahlen (z. B. zur quantitativen Darstellung der Häufigkeit von Behandlungskomplikationen – konkrete Datentypen; Kap. 2.2.1), die mit unterschiedlichen mathematisch-statistischen Verfahren analysiert werden können. Wenn Datensätze mehrerer Einrichtungen zusammengeführt werden, können Mittelwerte, Varianzen, oder Grenzwerte oder Abweichungen errechnet oder Auffälligkeiten und (un)typische Muster ermittelt (ein qualitätsbezogenes Profiling einzelner Einrichtungen), aber auch direkte Vergleiche unterschiedlicher Einrichtungen möglich werden (qualitätsbezogene Rankings/Scorings von Einrichtungen).

Benchmarkanalysen spielen bei der externen Qualitätssicherung im Rahmen der Selbstverwaltung des Gesundheitssystems eine zunehmend wichtige Rolle. Der Gemeinsame Bundesausschuss ist für die Definition unterschiedlicher Qualitätsdimensionen und entsprechender Indikatoren/Kennziffern verantwortlich. Er kann freiwillige Initiativen zulassen (z. B. Analysen anhand von Traumaregisterdaten für die Notfall- und Intensivmedizin; Kap. 4.1.4) und medizinische Einrichtungen zur Teilnahme an QS-Verfahren verpflichten, sofern Analysekonzepte entwickelt und in Fachgremien abgestimmt und akzeptiert wurden. Inzwischen gibt es diverse, vor allem sektorspezifische QS-Verfahren. Eine besondere Herausforderung sind jedoch die Konzeption sektorübergreifender QS-Verfahren und die Festlegung der jeweiligen Datenmodelle (ausführlicher Döbler/Follert 2021; Mühr 2018). Vor allem Behandlungsprozesse, die unter Public-Health-Gesichtspunkten höchstrelevant sind (z. B. die Versorgung chronische erkrankter und/oder multimorbider Personen), sind fachlich oft komplex, weil mehrere medizinische Einrichtungen an der Behandlung beteiligt sind. Zudem sind etliche relevante Aspekte (wie z. B. Kooperation unterschiedlicher an der Behandlung beteiligter Einrichtungen) aufwendig in der Erfassung, teilweise fehlen Konzepte, wie diese mittels Kennziffern dargestellt werden können. Im Gegenzug sind Kennziffern, die aus der Behandlungsdokumentation und/oder aus Leistungsabrechnungsdaten relativ leicht abgeleitet werden können, zwar nötig, erfassen jedoch nicht immer die relevantesten Aspekte eines Sachverhalts (Döbler/Follert 2021, S. 244). Insbesondere patientenzentrierte Kennziffern (z. B. ausreichend Zeit für und gut verständliche Arzt-Patienten-Kommunikation) werden national bei QS-Verfahren kaum berücksichtigt (Mühr 2018, S. 7 ff.).

Aus pragmatischen Gründen bauen QS-Verfahren bisher überwiegend auf Daten auf, die medizinische Einrichtungen aus ihren Behandlungsdaten ableiten. Ergänzend werden mitunter Registerdaten verwendet (z. B. aus klinischen Krebsregistern; Kap. 4.1.4). Bereits diese Datenzusammenstellungen sind aufgrund der geringen Standardisierung der primären Behandlungsdokumentationen mit erheblichem Aufwand verbunden. Medizinische Einrichtungen übermitteln ihre QS-Datensätze an die mit den externen QS-Analysen betrauten Institutionen (Kassenärztliche Vereinigungen im ambulanten Bereich [Kap. 5.3]; Landesgeschäftsstellen für Qualitätssicherung bei den Landeskrankenhausgesellschaften im stationären Bereich). Bei sektorübergreifenden QS-Analysen leiten diese Institutionen die Datensätze an das Institut für Qualitätssicherung und Transparenz im Gesundheitswesen (IQTIG) weiter. Alle haben separate Stellen für die Datenannahme, die periodenübergreifende Pseudonymisierung sowie für die Verwaltung der zunehmend großen QS-Datenbestände, bei den beteiligten Institutionen in einem geschützten Datenraum bleiben, nur für QS-Analysen verwendet und nicht mit anderen Daten verknüpft werden dürfen (§ 299 SGB V). Standardisierte einrichtungsspezifische Auswertungen können über die Vertrauens-/Pseudonymisierungsstellen den jeweiligen medizinische Einrichtungen zugeordnet und übermittelt werden (individuelles Feedback). Ergebnisse verpflichtender QS-Verfahren werden im Rahmen der Selbstverwaltung diskutiert und können sich inzwischen auch auf die Vergütung der medizinischen Einrichtungen auswirken (Zu- oder Abschläge) und rechtliche Konsequenzen haben (bis zum Entzug von Genehmigungen).

Bisher sind nur Krankenhäuser verpflichtet, jährlich strukturierte Qualitätsberichte über ihre Homepage zu veröffentlichen (§ 136b SGB V) und maschinenlesbare Versionen an den G-BA zu übermitteln. Dafür führen sie die Ergebnisse aller durchgeführten Maßnahmen zum Qualitätsmanagement und zur Qualitätssicherung zusammen. Arztpraxen müssen keine praxisspezifischen Qualitätsberichte veröffentlichen. Bisher veröffentlichen die Kassenärztlichen Vereinigungen nur aggregierte Berichte des ambulanten Bereichs.

Es gibt unterschiedliche Auffassungen, inwiefern der Aufwand für die Datenzusammenstellung und der Nutzen aus den QS-Analysen in einem angemessenen Verhältnis stehen. Insbesondere im ambulanten Bereich wird teilweise massive Kritik geäußert: Die grundsätzlich gute Idee der sektorübergreifenden QS-Verfahren hätte sich verselbstständigt; man habe ein für medizinische Einrichtungen hyperkomplexes und aufwendiges System geschaffen, bei dem der unmittelbare Nutzen für Patient/innen oftmals unklar sei.[123] Besondere Hürden für die Realisierung von QS-Verfahren sehen Döbler/Follert (2021, S. 240 ff.) nach wie vor in der geringen Interoperabilität medizinischer Behandlungsdaten. Dadurch könnten insbesondere medizinische Leistungen, die Patient/innen in

123 www.kbv.de/html/sqs.php (2.11.2021)

unterschiedlichen Einrichtungen erhalten, nicht immer eindeutig dem primär zu untersuchenden Sachverhalt zugeordnet werden. Komplexe Datenmodelle seien unausweichlich und in Folge sei auch die Entwicklung und Implementierung der Analysekonzepte aufwendig. Döbler/Follert sprechen sich dafür aus, ergänzend auch leistungsdatennutzende QS-Ansätze weiterzuentwickeln. Diese Daten haben zwar keine große medizinisch Detailgenauigkeit, dafür seien die Analysekonzepte aber auch nicht so kompliziert.

4.4.3 Daten für die amtliche Statistik

Seit Jahrzenten verpflichtet die Krankenhausstatistik-Verordnung (KHStatV)[124] alle Krankenhäuser in Deutschland (einschließlich Vorsorge- und Rehabilitationseinrichtungen, die teilweise von Renten- und Unfallversicherungen getragen werden) jährlich einen Datensatz mit drei Teilen (Grunddaten, Diagnosestatistik, Kosten [§ 3 KHStatV]) zusammenzustellen und an die zuständigen Statistischen Landesämter zu übermitteln. Diese Daten werden aus den medizinischen und administrativen Daten auf Einzelfallebene mit den jeweiligen KIS ermittelt, sind jedoch Zusammenfassungen (Gruppenwerte, von denen nicht auf ein einzelnes Subjekt geschlossen werden kann). Die Bezugseinheit ist das einzelne Krankenhaus. Statistische Landesämter als datenempfangende Stellen sind für die Prüfung dieser Datensätze verantwortlich, ggf. berichtigen sie diese. Diese Daten werden ebenfalls bundesweit zusammengeführt und gelten als absolut anonymisiert. Sie werden nach etwa 12 Monaten veröffentlicht (Fachserie 12, Reihe 6.1 bis 6.3). Etwa 18 Monate nach dem Berichtsjahr sind kontrolliere Datenabfragen über die Forschungsdatenzentren der statistischen Ämter möglich (Kap. 3.3.3) (Schubert et al. 2014, S. 49 ff.).

4.4.4 Gesamteinschätzung der analytischen Potenziale administrativer Daten

Es steht außer Frage, dass die administrativen Daten, die medizinische Einrichtungen regelmäßig zusammenstellen müssen, für vielfältige analytische Fragestellungen auf unterschiedlichen Ebenen (innerbetrieblich wie gesundheitssystemisch) wichtig sind. Etliche Analysen sind als öffentliche Aufgaben explizit gesetzlich definiert und mit diesen Aufgaben spezifische Einrichtungen betraut worden (z. B. Analysen zur Qualitätssicherung). Für die Bewertung des analytischen Potenzials unterschiedlicher Datenbestände sind viele Faktoren relevant, darunter der aufgabenspezifische Informationsgehalt), die Aktualität und

124 Verordnung über die Bundesstatistik für Krankenhäuser (Krankenhausstatistik-Verordnung – KHStatV)

Richtigkeit der Daten, die semantisch und syntaktisch standardisierte Darstellung sowie die Zugänglichkeit. Ein besonderes analytisches Potenzial wird den Leistungsabrechnungsdaten unterstellt. Sie werden längst nicht nur zur Leistungsabrechnung mit den jeweiligen Kostenträgern genutzt. Auch für vielfältige gesundheitssystemische Planungs- und Entwicklungsaufgaben, wie z. B. der Fortschreibung der Abrechnungsverfahren (Kap. 5.2) aber auch zu Forschungszwecken, können sie verwendet werden. Die Versorgungsforschung des Gesundheitssystems basiert in erheblichem Maße auf Leistungsdatenbeständen.

Die von medizinischen Einrichtungen zusammenzustellenden administrativen Datensätze können zudem auch zur Optimierung innerbetrieblicher Geschäftsprozesse genutzt. Wegen des hohen Kostendrucks in medizinischen Einrichtungen gewinnen betriebswirtschaftliche Analysen zur Erschließung möglicher Wirtschaftlichkeitsreserven seit Jahren an Bedeutung. PIS-/KIS-Hersteller bieten zunehmend spezifische Module an, mit denen u. a. die Auslastung einrichtungsinterner Kapazitäten, Aufwandskomponenten (z. B. Kosten für Personal, Material und Behandlung) und Ergebniskomponenten (z. B. von erfolgreichen Behandlungen bis zu Komplikationen, Rehospitalisierungen) ermittelt und in Relation gesetzt werden können. Derartige Business-Intelligence-Analysen könnten ebenfalls als Data-Mining interpretiert werden. Weichert (2018, S. 46) weist darauf hin, dass es bei betriebswirtschaftlichen Optimierungen regelmäßig einen Zielkonflikt zwischen innerbetrieblicher Kostensenkung und optimaler Behandlung gibt und dass, sofern Algorithmen für derartige Berechnungen eingesetzt werden, darauf zu achten sei, dass definierte Behandlungsstandards berücksichtigt und nicht unterschritten werden. Eine substanzielle Auseinandersetzung mit den Chancen und Risiken von Data-Mining-Ansätzen in betriebswirtschaftlichen Prozessen medizinischer Einrichtungen ist im Rahmen dieser Überblicksarbeit nicht möglich. Damit werden grundsätzliche Probleme des wirtschaftlichen Einflusses auf medizinische Entscheidungen angesprochen (Stichwort Ökonomisierung der Medizin), deren Folgedimensionen substanziell nur in eigenständigen Untersuchungen thematisiert werden können.

5 Data-Mining im Gesundheitssystem

Vielfältige gesundheitssystemische Aktivitäten sind als Aufgaben im öffentlichen Interesse gesetzlich definiert und werden datenanalytisch realisiert. Dafür werden medizinische Einrichtungen verpflichtet, unterschiedliche administrative Datensätze zusammenzustellen (Kap. 4.4) und an gesundheitssystemische Einrichtungen/Organe/datenverarbeitende Stellen weiterzuleiten. Dem Untersuchungsauftrag entsprechend werden in diesem Kapitel datenanalytisch relevante Institutionen des öffentlichen Gesundheitssystems mit ihren gesetzlich definierten Aufgaben, ihren Datenbeständen und den sich dadurch eröffnenden Data-Mining-Möglichkeiten und -Grenzen dargestellt.

Der Fokus der nachfolgenden Darstellung liegt auf den Leistungsabrechnungsdaten, die von den unterschiedlichen medizinischen Einrichtungen zusammengestellt werden und in mehreren Etappen zusammengeführt, analysiert, weitergeleitet und weiterverwendet werden können. Aus rechtlicher Sicht werden diese bei den gesundheitssystemischen Einrichtungen gespeicherten Leistungsdaten auf Einzelfallebene als Sozialdaten bezeichnet und geschützt (Kap. 5.1). Die nachfolgende Kapitelstruktur orientiert sich an den unterschiedlichen Bereichen des nationalen Gesundheitssystems und deren Datenflüsse. Im stationären Bereich rechnen medizinische Einrichtungen ihre Leistung zwar direkt mit den einzelnen Krankenkassen ab, jedoch laufen beim Institut für Entgeltsysteme im Krankenhaus (InEK) umfangreiche Leistungsabrechnungsdatenbestände zusammen, um mit komplexen Datenanalysen das Leistungsabrechnungsverfahren für den gesamten stationären Bereich regelmäßig fortzuschreiben (Kap. 5.2). Im ambulanten Bereich fungieren die Kassenärztliche Vereinigungen (Kap. 5.3) und Apothekenrechenzentren (Kap. 5.4) als intermediäre Einrichtungen der Leistungsabrechnung mit den Krankenkassen. Die Leistungsabrechnungsdaten der unterschiedlichen Bereiche laufen bei den Krankenkassen zusammen (Kap. 5.5). Vor allem große Krankenkassen, die Leistungsabrechnungsdaten von vielen medizinischen Einrichtungen und vielen Versicherten verwalten, haben besondere Data-Mining-Möglichkeiten. Anhand unterschiedlicher Anwendungsbeispiele werden diese Möglichkeiten aber auch die Herausforderungen und Grenzen der sekundären Nutzung von Leistungsabrechnungsdaten im Rahmen von Data-Mining-Prozessen veranschaulicht.

5.1 Sozialdaten: Rechtsrahmen der Verarbeitung

Datensätze zur Leistungsabrechnung haben einerseits Patienten- und Gesundheitsbezüge. In der DSGVO gelten sie als personenbezogene Daten besonderer Kategorie (Kasten 3.6). In der nationalen Sozialgesetzgebung werden sie allge-

meiner als Sozialdaten bezeichnet (§ 67 SGB X). Andererseits haben diese Datensätze auch Bezüge zu medizinischen Einrichtungen (teilweise sogar zu behandelnden Ärzte/innen) sowie zu den jeweiligen Krankenkassen als Leistungsträger (Abb. 4.4). Da medizinische Einrichtungen und Krankenkassen als eigenständige Wirtschaftsbetriebe agieren, können zumindest Teile der Leistungsdatensätze als betriebs- oder geschäftsbezogene Daten mit Geheimnischarakter aufgefasst werden (sie sind Sozialdaten gleichgestellt [§ 35 Abs. 4 SGB I; § 67 SGB X]). Der Umgang mit diesen Daten wird grundsätzlich im zweiten Kapitel des SGB X (Schutz von Sozialdaten) in Kombination mit § 35 SGB I (Sozialgeheimnis für datenverarbeitende Stellen) definiert. Laut Bundesbeauftragtem für Datenschutz und Informationsfreiheit ist das Sozialgeheimnis der ärztlichen Schweigepflicht weitgehend gleichrangig.[125]

Einrichtungen, die gesetzlich definierte Aufgaben im Rahmen der sozialen Sicherung ausführen, sind grundsätzlich zum Datenschutz und zur Geheimhaltung verpflichtet und dürfen Sozialdaten nur im Rahmen ihrer Befugnis verarbeiten (§ 35 SGB I). Datenvermeidung und Datensparsamkeit werden inzwischen nicht mehr als Grundprinzipien des Umgangs mit Sozialdaten genannt. Unterschiedliche Einrichtungen der sozialen Sicherung verwalten jeweils eigenständige Stammdatenregister und Fachdatenbestände. Direktidentifizierende Angaben dürfen inzwischen über ein zentralisiertes Verfahren aktualisiert und abgeglichen werden. Für den Bereich der gesetzlichen Krankenversicherung konkretisiert das SGB V die Aufgaben unterschiedlicher öffentlicher Einrichtungen und deren Befugnis zur Analyse von Sozialdaten.

Auch jenseits gesetzlich definierter Aufgaben ist sowohl die einrichtungsinterne Datennutzung für bestimmte Forschungs- oder Planungsvorhaben als auch die Datenübermittlung an Dritte für entsprechende Vorhaben im Sozialleistungsbereich zulässig, soweit diese Daten dafür erforderlich sind, schutzwürdige Interessen Betroffener nicht beeinträchtigt werden oder das öffentliche Interesse an der Forschung oder Planung dem Geheimhaltungsinteresse Betroffener erheblich überwiegt. Dafür müssen identifizierende Merkmalsbereiche (auch als Stammdaten bezeichnet) grundsätzlich von Fachdaten getrennt gespeichert und letztere sobald dies für den Forschungszweck möglich ist, anonymisiert werden. Dafür muss das Vorhaben einschließlich Datenschutzkonzept von der jeweiligen Aufsichtsbehörde für bundesweit agierende gesetzliche Krankenkassen vorab genehmigt werden. Soweit zumutbar muss eine Einwilligung bei betroffenen Personen eingeholt werden (§ 75 SGB X). Fraglich ist, wie oft Zumutbarkeit bisher tatsächlich attestiert wurde. Da es jenseits postalischer oder persönlicher Anfragen bisher gar keine Möglichkeiten gab, um Einwilligungen einzuholen, dürfte vor allem bei umfangreichen Datenanalysen, in die

125 www.bfdi.bund.de/DE/Buerger/Inhalte/GesundheitSoziales/IhreRechte/Sozialgeheimnis.html (2.11.2021)

Daten von vielen Personen einflossen, diese Einwilligungseinholung regelmäßig als unzumutbar eingestuft worden sein. Inzwischen sind auch elektronische Einwilligung in Datenanalysen zu Forschungszwecken rechtlich zulässig (§ 67b Abs. 3 SGB X). Fraglich ist wann entsprechende Einwilligungsmanagementsysteme technisch einsatzbereit sind.

Datenmissbrauch ist die nichtbefugte Verarbeitung und/oder Zugänglichmachung von Sozialdaten. Vorsätzliche oder fahrlässige Handlungen gelten als Ordnungswidrigkeiten (§ 85 SGB X), erst die vorsätzliche Zugänglichmachung gegen Entgelt mit Bereicherung oder Schädigung ist eine Straftat (§ 85a SGB X). Diese Detailregelungen des SGB X zeigen, dass es doch gewisse rechtliche Unterschiede zur ärztlichen Schweigepflicht gibt.

In diesem rechtlichen Rahmen werden Data-Mining-Aktivitäten im Kontext der sozialen Sicherung möglich (Kap. 5.3 u. 5.5).

5.2 Das Institut für das Entgeltsystem im Krankenhaus: Daten und Analytik

Das 2001 gegründete und als GmbH organisierte Institut für das Entgeltsystem im Krankenhaus (InEK)[126] unterstützt die Krankenhäuser und Krankenkassen bei der Leistungsabrechnung. Dafür schreibt das InEK das nationale Vergütungssystem stationär erbrachter Behandlungsleistungen – auch als *System of German Diagnosis Related Groups* (G-DRG-System) bezeichnet – regelmäßig datenbasiert fort und aktualisiert den Grouper-Algorithmus, durch den nahezu alle stationär behandelten Personen[127] als Behandlungsfälle anhand definierter Merkmale diagnosebezogenen Fallgruppen zugeordnet werden und die gruppenspezifische pauschalierte Vergütung berechnet wird. Die Fortschreibung des G-DRG-Systems basiert auf einem jährlich neu zusammenzustellenden, spezifischen nationalen Datenbestand und komplexen datenanalytischen Prozessen.

5.2.1 Die stationäre Leistungsvergütung als lernendes System

In Deutschland wurden bis Anfang der 1990er Jahre alle stationären Behandlungsleistungen anhand von allgemeinen Pflegetagessätzen vergütet. Dieses Verfahren war mit einem moderaten Dokumentations- und Abrechnungsaufwand verbunden, galt bezüglich der jeweiligen stationär erbrachten Leistungen jedoch als intransparent und setzte Fehlanreize (u. a. hohe Verweildauer von Pa-

126 Die InEK GmbH hat die Deutsche Krankenhausgesellschaft, den GKV-Spitzenverband und den Verband der privaten Krankenversicherungen als Gesellschafter.

127 Psychiatrische und psychosomatische Leistungen werden in Deutschland nach wie vor durch ein pauschaliertes Entgeltsystem abgerechnet (Bundesregierung 2016c).

tient/innen in Krankenhäusern, tendenziell innovationshemmend, fehlende Anreize zum effizienten Wirtschaften). Um diese strukturellen Probleme zu überwinden, wurde die Leistungsabrechnung im stationären Bereich in den 1990er Jahren schrittweise auf differenziertere Pflegesätze und dann Fallpauschalen umgestellt, die sich an den Diagnosen der Patient/innen orientieren, zuerst freiwillig in einigen Krankenhäusern für eine begrenzte Menge an Indikationen (z. B. Blinddarmoperationen). Die gesammelten Erfahrungen verstärkten den gesundheitspolitischen Willen, die Vergütung von Krankenhausleistungen möglichst vollständig auf Fallpauschalen umzustellen – ein zu dieser Zeit weltweit einzigartiger Ansatz, der damals wie heute kontrovers diskutiert wird (dazu und im Folgenden Beivers/Emde 2020; Schepers et al. 2015, S. 72 ff.). Mit dem gewählten Ansatz sollten das Leistungsgeschehen im Krankenhaus transparenter, der Ressourceneinsatz bedarfsgerechter und effizienter sowie die erbrachten Behandlungsleistungen angemessener honoriert werden. Durch die Erschließung von Wirtschaftlichkeitsreserven sollten leistungsstarke Krankenhäuser einen Wettbewerbsvorteil erhalten und zudem die Ausgaben der GKV stabilisiert werden. Man versprach sich davon auch, dass strukturelle Probleme des stationären Bereichs überwunden werden könnten (u. a. Senkung der Verweildauer, Abbau von Überkapazitäten).

Die politischen Weichen für das neue Vergütungssystem wurden 2000 mit dem GKV-Gesundheitsreformgesetz gestellt. In Fachgremien wurde u. a. festgelegt, welche Parameter für die Leistungsabrechnung herangezogen werden sollen (in erster Linie Diagnosen und Behandlungsleistungen, für die es bereits Klassifikationen gab, mit denen die Angaben einheitlich codiert werden konnten; Kasten 4.3). Durch die Überarbeitung des Krankenhausfinanzierungsgesetzes (KHG), des Fallpauschalengesetzes (FPG) und des Krankenhausentgeltgesetzes (KHEntgG) konnte das neue Vergütungssystem ab 2003 flächendeckend eingeführt werden (das Leistungsprinzip ersetzte das Selbstkostendeckungsprinzip).

Kern des neuen Vergütungssystems ist ein nationaler Fallpauschalenkatalog. Er enthielt beim Start 2003 knapp 660 diagnoseorientierte Fallgruppen mit spezifischen Basisfallwerten, Gewichtungsfaktoren entsprechend dem jeweiligen Schweregrad eines Behandlungsfalls (sogenannte Bewertungsrelationen) sowie Zu- oder Abschlägen je nach Länge des Krankenhausaufenthalts. Der Katalog wird jährlich fortgeschrieben und an die sich ändernden medizinischen Möglichkeiten und die damit verbundenen Kosten angepasst. Für die Leistungsabrechnung wird jeder stationäre Behandlungsfall anhand von insgesamt 1.500 Diagnose- und 26.000 Prozedurencodes beschrieben, anhand der codierten Haupt- und Nebendiagnosen und medizinischen Behandlungen einer von inzwischen 1.300 Fallgruppen zugeordnet, nach der Erkrankungsschwere gewichtet und je nach Behandlungsverlauf die Vergütung der gesamten stationä-

ren Behandlung der Erkrankten ggf. mit Zu- oder Abschlägen ermittelt. Der Katalog und die Zuordnungsregeln werden anhand spezifischer Fachkenntnisse, umfangreicher jährlich neu zusammenzustellender Datensätze und komplexer datenanalytischer Verfahren fortgeschrieben. Wegen der jährlichen Fortschreibung wird das G-DRG-System auch als *lernendes System* bezeichnet (InEK 2020, S. 43). Die jährliche G-DRG-Systemfortschreibung gilt als datenanalytisches Großprojekt. Allein am InEK sind 50 Personen mit der Realisierung beschäftigt. Ein erheblicher Teil der notwendigen Arbeiten findet jedoch in den Krankenhäusern statt, weit dort vielfältige Einträge in den arztgeführten Primärakten codiert und die erforderlichen Datensätze sowohl für die Leistungsabrechnung als auch für die Systemfortschreibung erstellt werden.

5.2.2 Daten und deren Weiterverwendungsmöglichkeiten

Jedes mittels Fallpauschalen abrechnende Krankenhaus ist verpflichtet, im ersten Quartal definierte Vorjahresdatensätze zusammenzustellen. Diese haben zwei Teile: Ein Teil sind krankenhausbezogene *Strukturdaten* (u. a. [Intensiv-] Bettenzahl, Zu- oder Abschlagsvereinbarungen, Zahl der Arbeits- und Ausbildungsplätze [§ 21 Abs, 2 Nr. 1 KHEntgG]). Den zweiten Teil bilden alle einzelfallbezogenen *Leistungsdaten* (§ 21 Abs, 2 Nr. 2 KHEntgG; Abb. 4.4.1). Im Rahmen der Abrechnung entdeckte Fehler in den Leistungsdaten sollten bereits berichtigt worden sein, bevor die Krankenhäuser mithilfe spezieller KIS-Module die Struktur- und Leistungsdaten zusammenstellen und in einheitlichem Format an die zentrale Datenstelle übermitteln. Laut InEK (2020, S. 5) umfasst diese Vollerhebung ca. 1.450 Krankenhäuser mit ca. 22 Mio. voll- und teilstationären somatischen Behandlungsfällen pro Jahr.[128]

Für die Fortschreibung des Fallpauschalenkatalogs werden zudem *Kostendaten* für jede stationär erbrachte Leistung pro Behandlungsfall erhoben (auf Vollkostenbasis, auch administrative und Pflegekosten sind enthalten). Diese Kostendaten ermitteln jedoch nur sogenannte Kalkulationshäuser (20%ige Stichprobe). 2020 haben 282 Kalkulationshäuser ihre Ist-Kosten freiwillig auf vertraglicher Basis mit dem InEK aufgeschlüsselt und an die zentrale Datenstelle übermittelt (diese Tätigkeit wird separat vergütet).

Die Datenstelle prüft die Plausibilität aller pseudonymisierten einzelfallbezogenen Leistungs- und Kostendatensätze algorithmenbasiert. Auffälligkeiten werden den Krankenhäusern zurückgeschickt, dort manuell geprüft und ggf. berichtigt. Prüfung und Berichtigung müssen bis zum Ende des zweiten Quartals abgeschlossen sein. Dann noch bestehende unplausible Kostendatensätze werden für die DRG-Fortschreibung nicht berücksichtigt. Begleituntersuchungen

128 Dieser Datensatz stimmt nicht überein mit dem der amtlichen Krankenhausstatistik (Kap. 4.4.3), für den nur vollstationäre Behandlungsfälle (auch psychische und psychosomatische) berücksichtigt werden.

zur Fortschreibung des nationalen DRG-Systems weisen auf steigende Prüfaufkommen und Rechnungskorrekturquoten und damit verbunden auch einen steigenden Personalaufwand bei den Krankenhäusern und den Prüfeinrichtungen hin (Fürstenberg et al. 2013, S. S. XIV).

Nach der Prüfung werden die Datensätze patientenanonymisiert und zum einen an das InEK für die DRG-Systemfortschreibung und für die diesbezügliche Begleitforschung geschickt. Zum anderen werden Datenteile (keine Kostendaten) auch an die Vertragsparteien (Krankenhausgesellschaften und GKV- und PKV-Verbände) und die Statistischen Ämter auf Landes- und Bundesebene übermittelt. Letztere stellen patienten-, einrichtungs- und kassenanonymisierte Daten auf Antrag durch das Forschungsdatenzentrum als Scientific- oder als Public-Use-File bereit (Kap. 3.3.3) und publizieren aggregierte Daten (Fachserie 12, Reihe 6.4). Weitere Organe der Selbstverwaltung dürfen Struktur- und Leistungsdaten beim InEK anfordern (u. a. für Qualitätssicherheits- und Wirtschaftlichkeitsprüfungen oder für Fusionskontrollen). Dafür ist im Detail festgelegt, wer für welchen Zweck welche Daten anfordern und welche Art Profil von einzelnen Krankenhäusern oder Fachabteilungen erstellen darf (§ 21 Abs. 3a KHEntgG).

5.2.3 Die DRG-Systemfortschreibung – ein Data-Mining-Prozess?

Die Fortschreibung des G-DRG-Systems mit seinem Fallpauschalenkatalog und dem Grouper-Algorithmus ist eine Mischung aus wissensbasierter Erweiterung, gesundheitspolitischer Steuerung sowie datenbasierter Anpassung von Vergütungspauschalen und Gruppierungsregeln. Unterschiedliche medizinische Fachgesellschaften und Gremien werden in den Fortschreibungsprozess eingebunden. Eine rein datenbasierte Fortschreibung ist aufgrund der Komplexität der Versorgungsprozesse und der Absicht, neue medizinische Untersuchungs- und Behandlungsmethoden (z. B. Multigentests; Kap. 4.3.3) schrittweise in das Abrechnungssystem zu integrieren, wird gesundheitspolitisch bisher nicht angestrebt (kein selbstlernendes System) (Schepers et al. 2015, S. 72 ff.).

Anhand der neuen Datensätze etablierter Behandlungsverfahren und der Daten der zu integrierenden neuen Behandlungsmethoden werden Fallgruppen jährlich überarbeitet: Teilweise werden Fallgruppen neu zugeschnitten, Basiswerte, Gewichtungsfaktoren für bestimmte Diagnose- und Behandlungskonstellationen sowie Zu- und Abschläge neu berechnet und Gruppierungsregeln entsprechend aktualisiert. Am Ende dieses datenanalytischen Prozesses entsteht ein überarbeiteter Fallpauschalenkatalog mit allen Vergütungspositionen sowie ein aktualisierter Grouper-Algorithmus. Letzterer kann in die unterschiedlichen Krankenhausinformationssysteme integriert werden, um im folgenden Abrechnungszeitraum jeden neuen Behandlungsfall anhand definierter Kennziffern ei-

ner Fallgruppe zuzuordnen, den Gewichtungsfaktor, mögliche Zu- oder Abschläge zu ermitteln sowie die Vergütungspauschale für die Behandlung automatisiert zu errechnen (ausführlicher z. B. InEK 2020).

Aus analytischer Perspektive ließe sich diskutieren, inwiefern die Fortschreibung des nationalen DRG-Systems als Mischung aus wissensbasierter Erweiterung, gesundheitspolitischer Steuerung sowie datenbasierter Gruppierung und Vergütungsanpassung als Data-Mining aufgefasst werden kann oder streng methodisch nicht dazugehört, weil die Verfahren nicht gänzlich datenbasiert fortgeschrieben werden, sondern Fachkräfte den Prozess maßgeblich steuern. Faktisch entsprechen die Prozessschritte der Systemfortschreibung denen von Data-Mining-Prozessen im weiteren Sinn (Abb. 2.1). Daten aus unterschiedlichen Quellen werden zu einem großen Analysedatenbestand zusammengeführt. Mit diesen Daten werden statistische Kennziffern und Faktoren neu kalkuliert, Fallgruppen charakterisiert, Klassifikationsregeln angepasst und fortgeschrieben. Am Ende entstehen aktualisierte Fallpauschalenkataloge, Klassifikationsregeln und ein Algorithmus, mit dem diese Regeln auf neue Behandlungsfälle angewendet und Vergütungspauschalen automatisiert zugewiesen werden. Das InEK prüft die Richtigkeit aller Ergebnisse und liefert zahlreiche Kennziffern für die Begleitforschung, bewertet jedoch nicht den Nutzen der etablierten Verfahren.

5.2.4 Einschätzung

Intendierte und nichtintendierte Folgen des DRG-Systems werden seit Jahren kontrovers diskutiert (ausführlich z. B. Beivers/Emde 2020). Zur Disposition stehen zumeist das gesamte Vergütungskonzept (stationär erbrachte Leistungen vollständig mittels Fallpauschalen zu vergüten), die damit einhergehenden Veränderungen der Behandlungsprozesse in Krankenhäusern (Ökonomisierung der Medizin) und die Realisierung der mit dem DRG-System ursprünglich anvisierten Ziele (u. a. leistungsgerechtere und transparentere Vergütung, Sicherung der wirtschaftlichen medizinischen Versorgung, Effizienzsteigerungen, Begrenzung der GKV-Ausgaben, Stärkung des Wettbewerbs und Förderung des Strukturwandels, Überwindung von Fehlanreizen). Das diesbezügliche datenanalytische Vorgehen wird bei diesen Kontroversen eher weniger thematisiert.

Unstrittig ist die Einschätzung, dass durch die Umstellung auf leistungsdifferenziertere Vergütungen die stationäre Behandlungsdauer verkürzt, aufwendige Behandlungsformen und medizinische Innovationen schneller in die Regelversorgung überführt sowie Wirtschaftlichkeitsreserven erschlossen wurden. Die dafür notwendigen Dokumentations- und Abrechnungsverfahren führten zu mehr Transparenz in Bezug auf die medizinischen Prozesse für Kostenträger und Organe der Selbstverwaltung, gingen jedoch mit einem erheblichen

Dokumentations-, Prüf- und Abrechnungsaufwand sowie mit einer entsprechenden Ressourcenverschiebung einher, sowohl in den Krankenhäusern als auch bei den datenempfangenden Institutionen. Die entstehende große Datenbasis ermöglichte den Aufbau eines in Bezug auf den Detaillierungsgrad weltweit einzigartigen Fallpauschalensystems. Laut Beivers/Emde (2020, S. 7) ist der Aufbau dieses Systems zum großen Teil der Arbeit des InEK zu verdanken, das national wie international großes Ansehen für die gründlichen Kalkulationen genießt. Die nationalen Ambitionen zur hochdifferenzierten Abbildung des medizinischen Leistungsgeschehens führt zu einer kontinuierlich steigenden Systemkomplexität. Für Nicht-Expert/innen dürften die Gruppierungsregeln und die Vergütungsberechnungen nur schwer nachzuvollziehen sein (Beivers/Emde 2020, S. 8 nach Dieterich et al. 2019).

Einige Erwartungen, durch die Vergütungsumstellung strukturelle Probleme des stationären Sektors lösen zu können, erfüllten sich nicht. So konnten Überkapazitäten nicht wie erhofft über Wettbewerbsmechanismen abgebaut und Investitionen nicht gesichert werden. Problematisch bleibt zudem die Finanzierung bestimmter kostenintensiver Angebote, für die Krankenhäuser Sicherstellungsaufträge haben (z. B. umfangreiche intensivmedizinische Notfallversorgungskapazitäten), weil diese Finanzierung bei der Vergütungsumstellung nicht kostengerecht organisiert wurde. Einschränkend muss jedoch darauf hingewiesen werden, dass die Planung von Krankenhauskapazitäten sowie die Investitionsfinanzierung als Aufgaben der Daseinsvorsorge in den Verantwortungsbereich der Bundesländer fallen. Die Leistungsvergütung durch die GKV kann für derartige Probleme nicht verantwortlich gemacht werden.

Dazu kommen einige neue strukturelle Probleme, die mit dem G-DRG-System direkt in Verbindung gebracht werden: Es würde Anreize zur Mengenausweitung ärztlicher Leistungen setzen; Pflegeleistungen und die Qualität erbrachter Behandlungsleistungen würden nicht adäquat abgebildet werden; die sektorale Versorgung sei manifestiert worden, Anreize für intersektorale Behandlungsansätze fehlten; nach wie vor gäbe es Über-, Unter- oder Fehlversorgungen; die Ökonomisierung medizinischer Prozesse habe Arbeitsabläufe intensiviert, die Arbeitsbelastung kontinuierlich erhöht und würde medizinische Entscheidungen beeinflussen (ausführlicher z. B. Beivers/Emde 2020).

Aus der DRG-System-Perspektive eröffnen sich zwei Wege, um die genannten Probleme abzumildern oder gar zu überwinden. Eine Möglichkeit ist die Fortentwicklung und Nachjustierung des DRG-Systems, um weitere Aspekte bei der Entgeltberechnung zu berücksichtigen. Die andere Möglichkeit sind separate Vergütungsformen außerhalb des DRG-Systems. Für beide Wege sind politische Aushandlungen und Weichenstellungen erforderlich.

Es hat mehr als 10 Jahre gedauert, bis nach der Einführung des G-DRG-Systems die Berechnungsverfahren u. a. einige Gewichtungsfaktoren manuell verändert wurden, um Fehlversorgungen entgegenzuwirken. Es hat mehr als

15 Jahre gedauert, bis die sich kontinuierlich verschlechternde Situation in der Pflege zu weiteren Veränderungen im G-DRG-System führte. Dazu wurde in erheblichem Maße in die Kalkulationen des InEK eingegriffen: Die ursprünglich in den Fallpauschalen enthaltenen Pflegepersonalkosten wurden extrahiert und sämtliche Fallpauschalen, Bewertungsrelationen sowie Zu- und Abschläge ohne Pflegekostenanteile neu berechnet. Zudem wurden eigenständige Kennziffern für Pflegeleistungen im Fallpauschalenkatalog ausgewiesen, mit denen perspektivisch krankenhausindividuelle Pflegebudgets vereinbart und Pflegepersonalkosten unabhängig von den Fallpauschalen vergütet werden sollen.[129] Auch dieser Ansatz wird kontrovers diskutiert. Durch die mehrgleisige Vergütung unterschiedlicher Leistungsarten steigt die Komplexität der Entgeltberechnung für Krankenhäuser weiter. Bei der gewählten Erstattungsform für Pflegeleistungen würde man zudem in das Selbstkostendeckungsprinzip der 1990er Jahre zurückfallen und weitere Fehlanreize setzen (man würde Anreize schaffen, der Pflege möglichst viele Aufgaben zuzuordnen, zudem könnten Pflegekräfte aus anderen Segmenten ohne Selbstkostendeckung wie z. B. aus der Reha oder Altenpflege abwandern) (Beivers/Emde 2020, S. 17). Selbst wenn mit diesem Ansatz eine leistungsgerechtere Vergütung der Pflege in Krankenhäusern gelingen sollte, kann dies nur ein Element zur Bekämpfung des allgemeinen Pflegenotstands in Deutschland sein. Kritiker/innen warnen, dass das G-DRG-System seine steuernden Möglichkeiten zunehmend einbüßt und werfen die Frage auf, ob die Herauslösung der Pflegeleistungsvergütung der Beginn des Ausstiegs aus dem G-DRG-System sei (Beivers/Emde 2020, S. 16 f.).

Weitere Kritikpunkte am G-DRG-System sind zum einen die fehlende Berücksichtigung der Qualität stationär erbrachter Leistungen bei der Vergütungsberechnung. Dafür müssten zunächst tragfähige Konzepte entwickelt werden, wie die Qualität erbrachter Krankenhausleistungen zu bemessen sei und um welche Qualitätskennziffern das G-DRG-System und die Vergütungskalkulationen erweitert werden könnten. Zum anderen gibt es Forderungen, das G-DRG-System zu erweitern und auch sektorübergreifende Versorgungsleistungen zu berücksichtigen. Die derzeitigen unterschiedlichen Vergütungskonzepte im stationären und ambulanten Bereich werden maßgeblich für starke sektorale Trennung der medizinischen Versorgung verantwortlich gemacht, die zunehmend kritisiert wird. Auch diesbezügliche Weiterentwicklungen erfordern grundsätzlichere Veränderungen des G-DRG-Systems und würden dessen Komplexität weiter erhöhen.

Seit der Einführung des datenanalytischen Großprojekts »G-DRG-System« wird kontrovers diskutiert, ob die damit angestrebte bessere Ressourcenallokation erreicht wurde und ob die auf komplexen Datenerhebungen und -analysen

129 www.bundesgesundheitsministerium.de/krankenhausfinanzierung.html (10.11.2021)

aufbauende Ökonomisierung medizinischer Versorgungsprozesse einen gesellschaftlichen Mehrwert bringt. Durch die Betrachtung der dafür erforderlichen komplexen datenanalytischen Prozesse lässt sich beitragen, dass diese einen erheblichen Aufwand verursachen und neben den gewünschten Effekten auch zahlreiche Fehlanreize mit sich bringen. Dennoch wurde nie in Abrede gestellt, dass die medizinische Leistungsvergütung als Aufgabe im öffentlichen Interesse die deutschlandweite Zusammenführung und Verwendung besonders schützenswerter Patientendaten hinreichend begründet. Schepers et al. (2015, S. 81 f.) sehen das Verfahren sogar als nationales Referenzprojekt für die Erstellung und Zusammenführung großer hochsensibler Datenbestände und deren Nutzung im Gesundheitsbereich. Das Beispiel zeige, was politischer Gestaltungswille zu bewegen vermag.

5.3 Kassenärztliche Vereinigungen: Daten und Analytik

Kassenärztliche Vereinigungen (KVen) sind Körperschaften öffentlichen Rechts unter Rechtsaufsicht der jeweils für Gesundheit zuständigen Landesbehörde (§ 77 SGB V).[130] Es gibt jeweils eine pro Bundesland (Nordrhein-Westphalen hat zwei) sowie eine zentrale Bundesvereinigung (KBV), die unter Rechtsaufsicht des BMG steht. KVen sollen als zentrale Einrichtung aller niedergelassenen Vertragsärzt/innen die flächendeckende ambulante medizinische Versorgung für alle GKV-Mitglieder sicherstellen (§ 75 SGB V). Dafür steuern sie die Verteilung von Praxisniederlassungen, überwachen die Erfüllung vertragsärztlicher Pflichten, realisieren die ambulante Leistungsvergütung, prüfen und sichern Wirtschaftlichkeit und Qualität der ambulanten medizinischen Versorgung, beraten Vertragsärzt/innen und bekämpfen Fehlverhalten (ggf. können sie ihren Ärzte/innen verbindliche Auflagen machen oder Sanktionen aussprechen). Im Rahmen dieser *primären Aufgaben* werden bei den KVen erhebliche Datenflüsse gelenkt und große Datenbestände mit unterschiedlichen Bezügen zu Personen (sowohl Patient/innen als auch niedergelassene Ärzte/innen) und zu medizinischen Einrichtungen aufgebaut. KVen haben Sozial- und Betriebsgeheimnisse zu wahren und dürfen die Datenbestände nur für gesetzlich definierte Aufgaben in unterschiedlich pseudonymisierter oder anonymisierter Form verwenden (§ 285 Abs. 1, § 305a SGB V). Sie können für einzelne Aufgaben spezielle Einrichtungen gründen (§ 77 ff. SGB V). Datenanalytisch relevant sind

- das in den 1970er Jahren gegründete *Zentralinstitut für die Kassenärztliche Versorgung* (Zi) als Stiftung bürgerlichen Rechts (Träger sind die KVen

130 Niedergelassene Zahnärzte/innen haben eine weitgehend identische parallele Organisationsstruktur mit Kassenzahnärztlichen Vereinigung auf Landes- und Bundesebene. Im Bericht wird auf die Nennung der Parallelstrukturen von Zahnärzt/innen verzichtet.

und die KBV), in dem ca. 60 Personen an der Schnittstelle zwischen Wissenschaft und Praxis an Fragen der ambulanten Versorgungssituation arbeiten, wobei wirtschaftliche Fragestellungen und die Entwicklung von Marktstrukturen im Fokus stehen; dafür kann das Zi Daten der Kassenärztlichen Vereinigungen auf Antrag nutzen, aber auch eigene Datenerhebungen mit freiwilliger Teilnahme organisieren (z. B. Zi-Praxis-Panel)[131] und weitere Daten hinzuziehen (z. B. Sozial- und Wirtschaftsdaten, die über Forschungsdatenzentren und -infrastrukturen zugänglicher gemacht werden; Kap. 3.3.3);

- das in den 1990er Jahren gegründete *Ärztliche Zentrum für Qualität in der Medizin* (ÄZQ) als Gesellschaft bürgerlichen Rechts (Träger sind die KBV und die Bundesärztekammer), in dem ca. 25 Personen an Analysen und Stellungnahmen zu grundsätzlichen und methodischen Fragen von Qualität und Qualitätsförderung im ambulanten Bereich arbeiten, sowie
- das 2006 gegründete *Institut des Bewertungsausschusses* (InBA)[132] als Gesellschaft bürgerlichen Rechts (Träger sind die KBV und der GKV-Spitzenverband), in dem ca. 50 Personen an der kontinuierlichen Weiterentwicklung des ambulanten medizinischen Vergütungssystems der GKV arbeiten (ähnlich dem InEK für den stationären Bereich) (Schepers et al. 2015, S. 160).

5.3.1 Kennnummern, Register und Informationssysteme der Kassenärztlichen Vereinigungen

Alle ambulant tätigen Ärzte/innen, die Leistungen im Rahmen der GKV erbringen und abrechnen wollen, müssen Mitglied einer Kassenärztlichen Vereinigung sein (Kap. 3.1.1). Die KVen führen landesweite Arztregister und Betriebsstättenverzeichnisse. Die Registereintragungen werden immer umfangreicher. Ein Teil dieser Eintragungen ist gesetzlich definiert (u. a. Arztnummer, Facharztkennung einschließlich dessen Gültigkeitsbeginn und -ende, Titel, Namen, Geburtsdatum, Geschlecht, Praxisadresse [§ 293 Abs. 4 SGB V]). Mit Verweis auf die Sicherstellung der medizinischen Versorgung fordern die Kassenärztlichen Vereinigungen in den einzelnen Bundesländern in unterschiedlichem Maße weitere personenbezogene Daten besonderer Kategorie von ihnen Mitgliedern (u. a. Staatsangehörigkeit, Privatadresse und Kontaktdaten, Sprach-

131 ausführlicher unter www.zi.de (2.11.2021)

132 Das Institut des Bewertungsausschusses (InBA) entwickelt das ambulante medizinische Vergütungssystem der GKV, die einheitlichen Bewertungsmaßstäbe, kontinuierlich weiter.

kenntnisse, Schwerbehindertenstatus, Praxisausstattung, ggf. Barrierefreiheit).[133] Seit 2008 vergeben die KVen eigene eindeutige lebenslange Arztnummern (LANR) an ihre Ärzte/innen und Betriebsstättennummern (BSNR) für deren Praxen,[134] die bei jeder Leistungsabrechnung gegenüber der KVen anzugeben sind (KBV 2021). Damit gibt es im ambulanten Bereich zwei Kennnummernsysteme: zum einen die von den KVen vergebene und zum anderen die von der Ärztekammer herausgegebene bundeseinheitliche Arztnummer (BAN) (Kap. 4.1.1), ggf. ist auch das allgemeine Institutionenkennzeichen (IK) relevant, falls Praxen besondere Versorgungsleistungen direkt mit den Krankenkassen abrechnen (Kap. 4.1.2). Kritiker halten diese doppelten Nummernsysteme für ressourcenaufwendig und fehleranfällig. Diese Nummern sind von zentraler Bedeutung sowohl für die Arztpseudonymisierung administrativer Daten als auch für vielfältige Datenverknüpfungen. Aus analytischer Sicht werden Verknüpfungen unterschiedlicher Datenbestände mit unterschiedlichen Schlüsselnummern technisch zumindest erschwert, Ärzte/innen und Praxen möglicherweise weniger gläsern gegenüber Organen der Selbstverwaltung, die Zugang zu diesen Registerdaten haben.

Kassenärztliche Vereinigungen dürfen ihre Arztregisterdaten nur für definierte Aufgaben verwenden (§ 285 Abs. 1, § 305a SGB V). Dazu gehört auch die Weitergabe des gesetzlich definierten Datenanteils an die Kassenärztliche Bundesvereinigung, die diese Daten zu einem bundesweiten Arzt- und Betriebsstättenverzeichnis zusammenführt, dieses monatlich aktualisiert und sowohl KBV-intern im Rahmen definierter Aufgaben verwenden darf, als auch dem GKV-Spitzenverband und seinen Mitgliedern unentgeltlich übermittelt, die diese ebenfalls für definierte Aufgaben verwenden und nicht weitergegeben dürfen (§ 293 Abs. 4 SGB V). Die einzelnen KVen sowie Ärzte/innen bekommen nur bei berechtigtem Interesse eine begrenzte Einsicht in das bundesweite Verzeichnis.

Um die Register und Verzeichnisse nicht nur kontinuierlich zu aktualisieren, sondern auch im Rahmen der gesetzlichen Aufgaben umfangreich nutzen zu können, werden die Datenverwaltungssysteme zu Informationssystemen ausgebaut, in die datenanalytische Funktionalitäten für unterschiedliche Aufgaben bzw. Dienste für unterschiedliche Zielgruppen Schritt für Schritt integriert werden können. Ein Beispiel dafür ist der von den Kassenärztlichen Vereinigungen in Abstimmung mit den Ärztekammern entwickelte Suchdienst *Arztsuche in Deutschland* bzw. dessen mobile Version die *116117-App.*[135] Damit reagieren die Kassenärztlichen Vereinigungen und andere Ärztevertretungen auf die von

133 z. B. www.kvs-sachsen.de/fileadmin/data/kvs/img/Mitglieder/Arbeiten_als_Arzt/Arztregister/200217-Arztregisterantrag-Arzt-C.pdf (10.11.2021)

134 In der LANR ist die von der KBV eigenständig definierte Fachgruppenklassifikation enthalten. In der BSNR ist über den KV-Landes- oder Bezirksstellenschlüssel eine grobe räumliche Zuordnung möglich.

135 www.kbv.de/html/arztsuche.php; www.kbv.de/html/116117-app.php (10.11.2021)

unterschiedlichen externen Unternehmen Jahre früher gestarteten Arzt(bewertungs)portale. Die Kassenärztlichen Vereinigungen bewerben ihren Suchdienst mit den aktuellsten und validesten Daten zu Anschriften und fachlichen Qualifikationen (Facharzt- und Schwerpunktkompetenzen, Zusatzweiterbildungen, Kommunikationssprachen) aller in Deutschland niedergelassenen Vertragsärzt/innen und Vertragspsychotherapeut/innen. Der Dienst bzw. das Informationssystem verfügt bisher über eine regionale Suchfunktion und eine Funktion zur Buchung von Facharztterminen. Bewertungs- oder Scoringverfahren für Ärzte/innen und Praxen lehnen die Kassenärztlichen Vereinigungen im Gegensatz zu externen arztportalbetreibenden Unternehmen bisher ab. Nach eigenen Angaben soll die 116117-App perspektivisch um ein sprachgesteuertes Assistenzsystem erweitert werden, das Symptome/gesundheitliche Beschwerden aufnimmt und Hinweise gibt, ob eine Selbstbehandlung ausreicht, Ärzte/innen demnächst aufgesucht oder ein Rettungsdienst sofort angefordert werden sollte (Direct-to-Consumer-Systeme; Kap. 4.3.4). Derartige digitale Funktionen/ Werkzeuge können als Resultate von Data-Mining-Prozessen aufgefasst werden. Sie können das Aufgabenspektrum einzelner öffentlicher Einrichtungen erweitern sowie neuartige Geschäftsideen hervorbringen. Eine dezidierte Betrachtung der mit derartigen neuen Funktionalitäten von Informationsdiensten einhergehenden Chancen und Herausforderungen und ein Vergleich mit extern entwickelten Arzt(bewertungs)portalen sollte eigenständig thematisiert werden.

5.3.2 Daten zur Qualitätssicherung und Data-Mining-Potenziale

Die KVen haben unterschiedliche Aufgaben bei der Realisierung der externen Qualitätssicherung (QS) im ambulanten Bereich. Bei datenbasierten *sektorspezifischen QS-Verfahren*, die u. a. im Rahmen von Disease-Management-Programmen und in derzeit 35 Leistungsbereichen (darunter die Durchführung von Mammografien oder der Umgang mit multiresistenten Infektionen) vereinbart sind,[136] übernehmen Kassenärztliche Vereinigungen die Koordination. Zum einen bauen sie die erforderlichen zentralen DS-Datenbestände auf. Dazu haben sie separate Stellen eingerichtet für die Datenannahme, Prüfung und Fehlerkorrektur eingehender spezifischer QS-Datensätze sowie für die Pseudonymisierung von patienten-, arzt- und praxisidentifizierenden Merkmalen (Vertrauensstelle). Zum anderen realisieren spezifische Analysestellen die standardisierten (Benchmark-)Analysen und erstellen einrichtungsspezifische Rückmeldungen und beraten ihre niedergelassenen Ärzte/innen ggf. diesbezüglich (Kap. 4.4.2). Zudem erstellen sie jährliche allgemeine Berichte zur Qualitätssicherung auf Meso- und Makroebene für den ambulanten Bereich. Bisher werden nur die allgemeinen Berichte veröffentlicht (Schepers et al. 2015, S. 170).

136 www.kbv.de/html/sqs.php (10.11.2021)

Weil insbesondere bei schwerwiegenden Erkrankungen oftmals sowohl ambulante als auch stationäre medizinische Behandlungsleistungen erforderlich sind, wurden 2010 zusätzlich s*ektorübergreifende* QS-Analysen verbindlich eingeführt. Bei diesen fungieren die KVen im Wesentlichen als Daten-annahme-, Prüf- und Vertrauensstelle (analog agieren im stationären Bereich die Landesgeschäftsstellen für Qualitätssicherung bei den Landeskrankenhausgesellschaften). Die zentralen Datenbestände und sektorübergreifenden Analysen werden jedoch beim explizit dafür gegründeten Institut für Qualitätssicherung und Transparenz im Gesundheitswesen (IQTiG) aufgebaut und realisiert.

Es gibt unterschiedliche Auffassungen inwiefern mit den derzeit etablierten Datenerhebungen die Qualität der medizinischen Versorgung tatsächlich erfasst werden kann, inwiefern der Aufwand in angemessenem Verhältnis zum Nutzen der Datenanalysen steht, welchen konkreten Mehrwert unterschiedliche komplexe Datenanalysekonzepte genau generieren und für welche Akteursgruppen handlungsrelevante Informationen generiert werden (sollen).

Die tendenziell weniger aufwendigen sektorspezifischen datenanalytischen Verfahren des ambulanten Bereichs sollen niedergelassene Ärzte/innen handlungsrelevante Informationen liefern und die Qualität der ambulanten Versorgung bei bestimmten Erkrankungen oder Behandlungsformen sichern. Kritiker/innen bemängeln vor allem deren intransparente Ergebnisse (Arztpraxen sind nicht zur Veröffentlichung ihrer Rückmeldungen verpflichtet). Bei den aufwendigeren sektorübergreifenden QS-Verfahren sind Kassenärztliche Vereinigungen der Meinung, dass sich die grundsätzlich gute datenanalytische Idee verselbstständigt habe und ein hyperkomplexes System geschaffen wurde, das für beteiligte Vertragsärzt/innen und KVen mit enormem Aufwand verbunden sei, wobei der Nutzen für Patient/innen unklar bleibe.[137] Bisher steht vor allem die akutmedizinische Versorgung im Zentrum derartiger Analysen. Patientenzentrierte Aspekte (z. B. Wartezeiten auf Facharzttermine, ausreichend Zeit für Arzt-Patienten-Gespräche, gut verständliche Information, Einbeziehen in Behandlungs- und Pflegeentscheidungen) werden bei gesundheitssystemischen QS-Analysen bis auf marginale Ausnahmen bisher nicht berücksichtigt (Mühr 2018). Diese starke medizinische und gesundheitssystemische Ausrichtung leistet u. a. externen Konzepten zur Praxis- und Arztbewertung Vorschub, die explizit auf patientenzentrierten Informationen aufbauen, ohne eine medizinisch hochdifferenzierte Darstellung und Analyse anzustreben. Ein weiteres konzeptionelles Problem bei sektorübergreifenden QS-Analysen ist die bisher fehlende praktische Entsprechung. Da es nach wie vor kaum sektorübergreifende Versorgungsformen gibt und Behandlungen in unterschiedlichen medizinischen Einrichtungen eigenständig dokumentiert werden, ist die Zuordnung unterschiedlicher stationär und ambulant erbrachter Behandlungsleistungen zu einem

137 www.kbv.de/html/sqs.php (10.11.2021)

Behandlungsfall aus den Dokumentationen der unterschiedlichen medizinischen Einrichtungen mit erheblichen Schwierigkeiten verbunden, mitunter auch nur begrenzt möglich (ausführlicher z. B. Döbler/Follert 2021). Die Forderung, bei sektorübergreifenden Analysen auch Präventions-, Reha- und Pflegeaktivitäten zu berücksichtigen, erhöht diese Schwierigkeiten. Dafür fehlen bisher sowohl Erfassungskonzepte als auch standardisierte Daten. Um die medizinische Versorgung insbesondere von Patient/innen mit erheblichen gesundheitlichen Beeinträchtigungen (z. B. chronisch Erkrankte oder multimorbide Personen) sektorübergreifend datenbasiert nachvollziehen und die diesbezügliche Qualität zu sichern, könnten einrichtungsübergreifende oder patientengeführte elektronische Akten hilfreich sein, sofern sie weitgehend vollständig geführt werden. Davon ist man im nationalen Gesundheitssystem derzeit noch weit entfernt.

Unterschiedliche Institutionen der Selbstverwaltung arbeiten kontinuierlich an der Weiterentwicklung datenanalytischer Konzepte zur Sicherung der Qualität der medizinischen Versorgung. Zum einen können sie ihre kontinuierlich größer werdenden expliziten QS-Datenbestände für sekundäre Datenanalysen einschließlich Data-Mining nutzen. Zudem gibt es seit Jahren Ansätze, anhand von Daten aus der Leistungsabrechnung Informationen zur Qualität der medizinischen Versorgung abzuleiten. Diese sind u. a. für Kassenärztliche Vereinigungen und Krankenkassen zugänglich, auch wenn KVen QS-Daten und Leistungsabrechnungsdaten nicht zusammenführen dürfen. Zur Sicherung der Behandlungsqualität bei Krebserkrankungen dürften außerdem die Krebsregisterdaten tendenziell wichtiger werden (Kap. 4.1.4), da diese perspektivisch Langzeitbeobachtungen und Längsschnittanalysen zulassen, wobei die kontinuierliche Dokumentation des Behandlungsverlaufs eigenständig geregelt und zudem finanziert wird. Diese unterschiedlichen Datenquellen bieten Möglichkeiten, u. a. mit Data-Mining-Ansätzen QS-Verfahren weiterzuentwickeln.

5.3.3 Leistungsdaten: Prüfung, Verwendung, Weiterleitung

Kassenärztliche Vereinigungen sind zentrale Intermediäre für die ambulante Leistungsabrechnung und -vergütung. Sie teilen die kollektivvertraglich mit dem GKV-Spitzenverband vereinbarte Gesamtvergütung für die ambulante medizinische Versorgung aller Versicherten auf alle teilnehmenden Vertragsärzt/innen auf (§ 87b SGB V). Dafür übermitteln letztere jeweils zum Quartalsende definierte, praxisintern geprüfte Leistungsdatensätze auf Patientenebene an die zuständige KV (Abb. 3.4).[138] Diese Leistungsdatensätze sind einerseits

138 Nicht enthalten sind medizinische Leistungen, die niedergelassene Ärzte im Rahmen spezieller Versorgungsformen direkt mit einzelnen Krankenkassen abrechnen, sowie Leistungen die sie als individuelle Gesundheitsleistungen den Patient/innen direkt in Rechnung stellen.

personenbezogene Daten besonderer Art, andererseits bilden sie die betrieblichen Prozesse der ambulanten Praxen in hohem Maße ab.

Für die Verarbeitung von Leistungsdaten haben die einzelnen KVen jeweils separate Datenannahme-, Vertrauens- und Analysestellen etabliert. In den Datenannahmestellen werden die eingehenden Leistungsdatensätze weitgehend automatisiert geprüft. Dafür werden erst die Datensätze um ausgewählte Angaben aus den selbst geführten Ärzteregistern ergänzt (u. a. Fachqualifikationen als Voraussetzung, um bestimmte Leistungen durchführen und abrechnen zu können). Dann werden sie mittels Auffälligkeitsprüfungen (Abweichungen bestimmter Prüfkriterien von Standardwerten [§ 296 SGB V]) und Zufälligkeitsprüfungen (Stichprobenüberprüfungen von Praxisdaten [§ 297 SGB V]) auf Plausibilität und Rechtmäßigkeit automatisiert geprüft. Bisher werden klassische regelbasierte Verfahren eingesetzt, die in Facharbeitsgruppen im Rahmen der Selbstverwaltung entwickelt und abgestimmt werden. Fehlerhafte Leistungsdaten werden gestrichen. Vertragsärzt/innen haben ein Beschwerderecht. Machen sie davon Gebrauch, erfolgt eine Einzelfallprüfung.

Nach der Prüfung pseudonymisieren die Vertrauensstellen die Leistungsdatensätze. Dann werden mit komplexen Berechnungsverfahren, in die u. a. Praxispauschalen, EBM-Kennziffern (quasi Pauschalen für ambulant erbrachte medizinische Leistungen), Fallwerte und -zahlen aber auch Vergütungsabstaffelungen im Rahmen von Mengenbegrenzen einfließen, die Quartalshonorare der Vertragsärzt/innen ermittelt (Schepers et al. 2015, S. 136). Diese Berechnung hat eine so große Komplexität erreicht, dass sie nur maschinell und automatisiert realisiert werden kann.

Nach der Prüfung und Vergütung werden die Leistungsdatensätze zum einen auf die unterschiedlichen Krankenkassen der Patient/innen aufgeteilt und die Teile den jeweiligen Kassen übermittelt (Kap. 5.5). Zum anderen wird der gesamte Leistungsdatenbestand von den Kassenärztlichen Vereinigungen für unterschiedliche gesetzlich definierte Zwecke weiterverwendet (u. a. die Fortschreibung des EBM-Katalogs, Wirtschaftlichkeits- und Qualitätsprüfungen sowie diesbezügliche Beratungen mit einzelner Vertragsärzt/innen). Dazu ist eine zeitliche Fortschreibung der einzelnen Leistungsdatensätze erforderlich (gesetzliche Grundlage: § 87 Abs. 3f SGB V). Patienten- und arztidentifizierende Merkmale werden mit schlüsselabhängigen Verfahren pseudonymisiert (Kap. 3.3.3) und dem Leistungsdatenbestand der einzelnen KVen und dem zentralen Bestand der KBV hinzugefügt. Im Data Warehouse der KBV wird der national größte Datenbestand ambulant erbrachter medizinischer Leistungen gespeichert. Er umfasst alle seit 2009 quartalsweise abgerechneten Leistungsdaten der 72 Mio. gesetzlich Versicherten (Schepers et al. 2015, S. 138).

Kassenärztliche Vereinigungen dürfen ihre Datenbestände für zeitlich befristete und vom Umfang begrenzte Forschungsvorhaben (insbesondere zur Ge-

winnung epidemiologischer Erkenntnisse sowie von Erkenntnissen über Zusammenhänge zwischen Erkrankungen und Arbeitsbedingungen oder über örtliche Krankheitsschwerpunkte) mit Erlaubnis der Aufsichtsbehörde leistungserbringer- oder fallbeziehbar selbst auswerten oder in anonymisierter Form über entsprechende Fristen hinaus aufbewahren (§ 287 SGB V). Für diese Analysen wurde auf Bundesebene u. a. ein strategisches Analyseteam eingerichtet. Das Team nimmt interne aber auch externe Analyseanfragen (z. B. vom BMG oder dem G-BA) entgegen, führt mit Erlaubnis der Aufsichtsbehörde die Analyse durch und meldet Ergebnisse ggf. zurück. Inzwischen scheint es deutlich über 100 interne und externe Themenanfragen pro Jahr zu geben (Tenckhoff 2017). Die Spanne reicht von mehr oder weniger standardisierten Analysen (z. B. zur Situation der ambulanten Versorgung z. B. urbanen und ländlichen Räumen) bis zur Entwicklung prognostischer Modelle (z. B. für Simulationsrechnungen, die die Entwicklung der ambulanten Versorgung prognostizieren sollen). Kassenärztliche Vereinigungen müssen eine Übersicht erstellen, wie sie ihre Bestandsdaten weiterverwenden (§ 286 SGB V).

5.3.4 Sekundärnutzung von Leistungsdaten: Data-Mining-Beispiel »Verbreitung multiresistenter Erreger«

Multiresistente Keime (Methicillin-resistenter Staphylococcus aureus – MRSA) sind gegen mehrere Antibiotika resistent und können oftmals nur schwer oder eingeschränkt behandelt werden. Aufgrund der begrenzten Behandlungsmöglichkeiten gelten sie als eine nationale Gesundheitsgefahr. Ähnlich wie beim Beispiel zur Choleraepidemie in London (Kap. 2.1) verweisen Schepers et al. (2015, S. 101 ff.) auf raumbezogene Data-Mining-Analysen, die die regionale Verbreitung multiresistenter Erreger und Hotspots der Verbreitung zeigen.

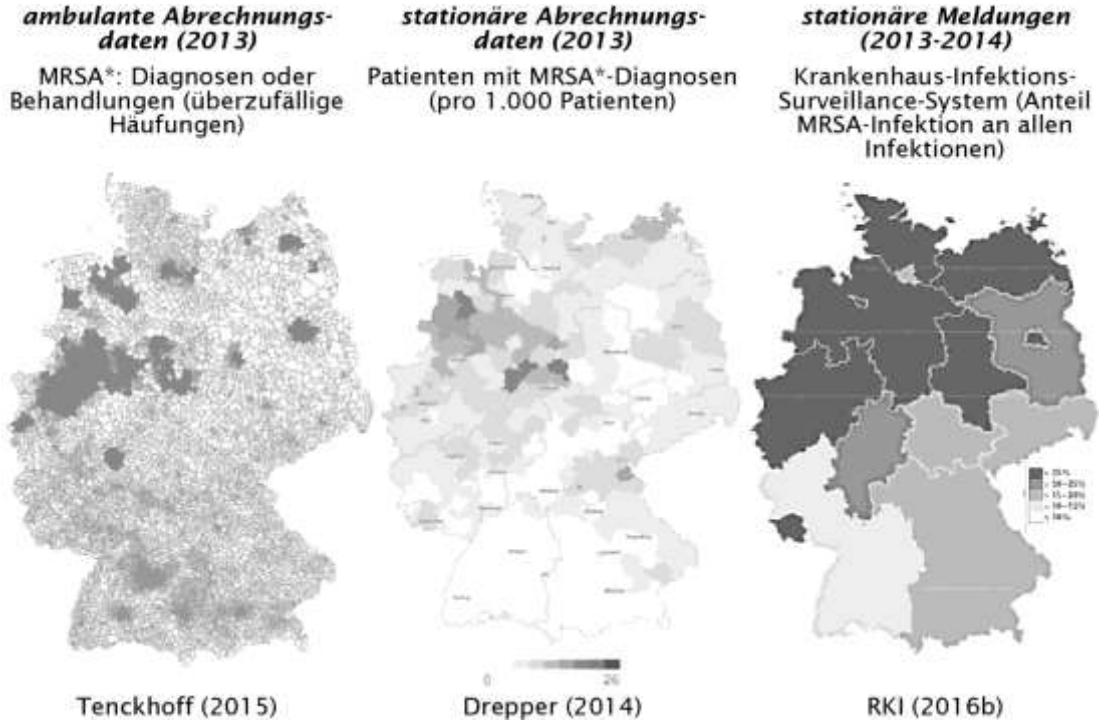
Tenckhoff (2015) (Abb. 5.1 links) hat aus den bei der KBV gehaltenen ambulanten Leistungsdaten von 2013 alle Fälle mit MRSA-diagnose- und MRSA-behandlungsbezogenen EBM-Codes[139] und deren Georeferenz (die auf 4 Stellen vergröberte PLZ des Wohnortes) extrahiert. Es wurden der nationale Mittelwert errechnet, regionale Abweichungen vom Mittelwert ermittelt und die Regionen anhand der Abweichungen gruppiert. Regionen, in denen überzufällig häufig MRSA-Diagnosen und/oder -Behandlungen abgerechnet wurden, wurden dunkel eingefärbt.

Drepper (2014) (Abb. 5.1 Mitte) hat aus den über das Statistische Bundesamt frei zugänglichen stationären Leistungsabrechnungsdaten (entsprechend § 21 KHEntgG; Kap. 4.2.2) alle 2013 in Krankenhäusern diagnostizierten MRSA-Patienten (ICD-Code: U80.0) extrahiert, die relativen Häufigkeiten (pro

139 EBM-Ziffern von 86770 – 86781 (für Aktualisierungen der Analyse von Tenckhoff (2015) müssen alle EBM-Ziffern umcodiert werden, weil der EBM-Katalog 2014 umcodiert wurde und sich alle MRSA-relevanten EBM-Ziffern änderten).

1.000 Patient/innen) ermittelt und diese ebenfalls über die verfügbare Georeferenz (Landkreis des Patientenwohnortes) grafisch dargestellt, wobei Landkreise mit hohen Häufigkeitswerten dunkel eingefärbt wurden.

Abb. 5.1 Regionale MRSA-Diagnosen in Deutschland (2013)



Quellen: Drepper 2014; RKI 2016; Schepers et al. 2015, S. 101 nach Tenckhoff 2015

Das RKI (2016) als nationales Public-Health-Institut erhält vielfältige Daten vor allem zur Verbreitung zu Infektionskrankheiten, darunter auch Daten die im Rahmen der Qualitätssicherung in Bezug auf MRSA-Krankenhausinfektionen erfasst werden. Für die Jahre 2013/14 wurden Daten von über 800 Intensivstationen und knapp 1.000 operativen Abteilungen analysiert und u. a. der Anteil der MRSA-Infektionen an allen festgestellten Krankenhausinfektionen ermittelt (Abb. 5.1 rechts).

Einschätzung und Bewertung

Zieht man die unterschiedlichen Schritte des Data-Mining-Prozesses als eine Art Bewertungsraster heran (Abb. 2.1), wird deutlich, dass die *Aufgabendefinition* bei allen drei Ansätzen weitgehend gleich ist: Anhand verfügbarer Daten sollen retrospektiv regionale Auffälligkeiten in Bezug auf MRSA-Infektionen ermittelt werden. Wo traten besonders viele, aber auch besonders wenige Fälle

auf? Die *Datenauswahl* dürfte durch die Zugangsmöglichkeiten der Datenanalyst/innen bestimmt worden sein. Der KBV-Analyst nutzte die ambulanten Leistungsabrechnungsdaten des KBV-Data-Warehouse (Tenckhoff 2015). Der investigative Journalist nutzte die allgemein zugänglichen stationären Leistungsabrechnungsdaten, die beim InEK zusammengestellt und über das statistische Bundesamt zugänglich gemacht werden (Drepper 2014). Das RKI nutzte die Daten zu labordiagnostisch bestätigten, meldepflichtigen Infektionen, die medizinische Einrichtungen an die Gesundheitsämter übermitteln und diese dann an das RKI weiterleiten. Die *Datenanalyse* basierte auf unterschiedlichen räumlichen Auflösungen bzw. geografischen Gebieten und methodischen Ansätzen: Tenckhoff ermittelte regionale Ausreißer , Drepper errechnete relative Häufigkeiten, das RKI setzte MRSA-Infektionen ins Verhältnis zu anderen Infektionsmeldungen. Zudem hatte jede Analyse eine unterschiedliche regionale Auflösung. Alle haben ihre geografischen Gebiete klassifiziert und dann anhand einer digitalen Karte dargestellt. Je größer die regionale Auflösung der Analysedaten, desto genauer können Häufungen verortet werden.

Aus gesundheitssystemischer oder -politischer Perspektive sind die Ergebnisse und deren weitere Verwendung relevant: Was folgt aus diesen Einzelaktivitäten und den Resultaten? Damit wird der Blick zuerst auf die fachlich-inhaltliche Diskussion gerichtet (*Wissenserweiterung, externe Prüfung*), ggf. kann auch die Verstetigung der Verfahren diskutiert werden (*Anwendung, Entscheidungsunterstützung*). Auch wenn die drei Analysen keine deckungsgleichen regionalen Muster hervorbrachten, liefern sie Hinweise/Signale zu regionalen Häufungen von MRSA-Infektionen im Nordwesten Deutschlands (Korrelationen), aber keine validen Belege (dafür gelten vor allem Abrechnungsdaten oftmals als zu fehleranfällig).[140] Die skizzierten retrospektiven Analysen können grundsätzlich keine Ursache-Wirkungs-Beziehungen aufzeigen (Kausalitäten). Sowohl die Daten (vor allem die hochaufgelösten) als auch die damit möglichen Analysen sind jedoch geeignet, um auf ein Problem hinzuweisen. Wie im historischen Beispiel der Choleraepidemie (Kap. 2.1) muss man Ursache-Wirkungs-Beziehungen auf andere Art und Weise nachgehen. Ziel der retrospektiven Datenanalyse ist die Darstellung der räumlichen Verteilung der Infektionen und die Ermittlung von MRSA-Infektions-Hotspots und die Darstellung räumlicher Strukturen, nicht mehr, aber auch nicht weniger. Darüberhinausgehende Fragen können mit diesen Analysen nicht beantwortet werden: beispielsweise die Frage nach der Infektionsquelle: Wurden MRSA-Infektionen in die jewei-

140 Niedergelassene Ärzte/innen können MRSA-Diagnose- und Behandlungsleistungen seit dem zweiten Quartal 2012 abrechnen, wenn sie eine zertifizierte MRSA-Zusatzqualifizierung haben und diese im Arztregister der KV eingetragen ist. Wie viele Ärzte/innen in welchen Regionen diese Qualifizierung haben und ob möglicherweise Teile Niedersachsens und Nordrhein-Westfalens Modelregionen für die MRSA-Zusatzqualifizierung sind, geht aus der Analyse nicht hervor.

lige Einrichtung mitgebracht oder dort erworben? Gibt es in nördlichen Regionen nur eine überdurchschnittliche MRSA-Aufmerksamkeit und wird dort nur genauer geprüft? Sind Keime in Regionen mit Intensivtierhaltung allgemein antibiotikaresistenter? Derartige vertiefte Auseinandersetzungen mit den Analyseergebnissen schließen an Data-Mining-Prozesse an und sind oftmals eingebettet in ohnehin laufende Aktivitäten zu einem bestimmten Sachverhalt. In Bezug auf die nationale Gesundheitsgefahr durch MRSA-Keime zählen dazu u. a.:

- Die sektorübergreifende *Ursachenforschung*: In diesem Rahmen untersucht z. B. das BfR (2014) Fragen zu einem möglichen Zusammenhang zwischen intensiver Nutztierhaltung und gehäuften MRSA-Besiedlungen beim Menschen und im interdisziplinären Forschungsverbund »HyReKA«[141] forschen Agrarwissenschaftler, Geografen, Ingenieure, Lebensmitteltechnologen, Ernährungswissenschaftler Biologen und Mediziner gemeinsam zur Rolle des Abwassers bei der Verbreitung antibiotikaresistenter Bakterien (One-Health-Ansatz).
- Maßnahmen zur *Qualitätssicherung* in der medizinischen Versorgung: 2016 trat die sektorspezifische QS-Vereinbarung zu MRSA im ambulanten Bereich in Kraft (KBV 2016). Zudem wird an einem sektorübergreifenden QS-Verfahren zu postoperativen Wundinfektionen (u. a. mit MRSA-Keimen) gearbeitet.[142] In diesem Rahmen müssen vielfältige Hygienemaßnahmen und Tests regelmäßig durchgeführt, Infektionen bekämpft und die Verfahren aufwendig dokumentiert werden. Dadurch sollte perspektivisch eine größere MRSA-spezifische Datenbasis generiert werden (die bisher nur im Rahmen der der Qualitätssicherungsverfahren begrenzt zugänglich sein wird).
- Kontinuierliches *Gesundheitsmonitoring*: Das RKI startete bereits 2007 das Antibiotika-Resistenz-Surveillance-Projekt. Es soll das eine repräsentative Datenbasis zur Antibiotikaresistenz in Deutschland aufbauen und unterschiedliche Datenbestände für diverse Analysen nutzen, um nicht nur regionale, sondern auch zeitliche und sektorspezifische Entwicklungen abzubilden (z. B. RKI 2016). In diesem Rahmen könnten auch die beispielhaft skizzierten geoanalytischen Verfahren verstetigt und als digitale Werkzeuge in MRSA-Informationssysteme integriert werden.

Die drei beispielhaft vorgestellten datenanalytischen Verfahren, die die räumliche Verteilung von MRSA-Infektionen anhand unterschiedlicher Analysedatensätze darstellen und mit unterschiedlichen Berechnungsverfahren Hotspots lokalisieren, wurden in einer Art Machbarkeitsstudie entwickelt (z. B. Tenckhoff

141 http://hyreka.net/ (10.11.2021)
142 Das 2017 eingeführte sektorübergreifende QS-Verfahren wurde 2020 aufgrund erheblichen Anpassungs- und Entwicklungsbedarfs sowie wegen unklarer Fragebogenformulierungen ausgesetzt (https://iqtig.org/qs-verfahren/qs-wi/; 10.11.2021).

2015). Dafür können in der Regel auch strukturgleiche alte Daten verwendet werden. Bevor die Verfahren z. B. im Rahmen des Gesundheitsmonitorings und/oder der Qualitätssicherung eingesetzt werden können (*Anwendung der analytischen Verfahren zur Entscheidungsunterstützung*), müssen unterschiedliche Detailfragen diskutiert und abgewogen werden: Welche der verfügbaren Datensätze sind inhaltlich und analysetechnisch am besten geeignet? Neben der Validität/Qualität ist auch die Aktualität und die räumliche Auflösung der Rohdaten relevant. Werden (Daten-)Schutzrechte eingehalten? Liegt bei öffentlichen Einrichtungen die Entwicklung von Informationsdiensten im Rahmen der gesetzlich definierten Aufgaben? Für welche Gesundheitsmonitorings eignet sich das Verfahren? Kann es in bestehende Informationssysteme als Zusatzfunktion integriert werden? Erreichen die Verfahren und die ermittelten Ergebnisse relevante Zielgruppen?

Es kann davon ausgegangen werden, dass derartige Fragen im Kontext von MRSA-Infektionen diskutiert und abgewogen wurden. Denn quartalsbezogene MRSA-Analysen durch die KBV und die Ergebnisberichterstattung an des BMG sind inzwischen gesetzlich definiert (§ 87 Abs. 2a Satz 3 f. SGB V).

5.3.5 Einschätzung

Kassenärztliche Vereinigungen sind als öffentliche Einrichtungen für vielfältige administrative Prozesse der ambulanten ärztlichen Versorgung die zentralen primären Datendrehscheiben. Im Rahmen ihres gesetzlich definierten Aufgabenspektrums bauen sie einzigartige Mikrodatenbestände auf (mit Bezügen zu Zeit, Raum, Gesundheit von Patient/innen, Behandlung, Ärzte/innen und deren Praxen sowie Krankenkassen) auf, die regelmäßig fortgeschrieben werden. Register- und Leistungsdaten sind bereichsspezifische Totalerhebungen, die relativ zeitnah für KV-interne sekundäre Analysen bereitstehen. Kassenärztliche Vereinigungen übermitteln definierte Teilbestände an andere Organe der Selbstverwaltung. Den Gesamtbestand dürfen nur sie (weiter)verwenden. Die Datenbestände oder Teile davon werden nicht über Forschungsdatenzentren oder Dateninfrastrukturen Dritten zugänglich gemacht. Kassenärztliche (Bundes-)Vereinigungen haben ein weitgehendes Verarbeitungsmonopol auf ihre Datenbestände.

Regelmäßigen komplexen Analysen sowohl zur Prüfung der Datensätze (sachliche und rechnerische Richtigkeit), zur Berechnung der Arzthonorare, zu Wirtschaftlichkeitsprüfungen und zur Qualitätssicherung wird auf gesetzlicher Ebene und durch vielfältige nachgeordnete Richtlinien ein klarer Rahmen vorgegeben, in dem die Kassenärztlichen Vereinigungen eigenverantwortlich agieren können. Zudem können sie ihre Datenbestände für eigene Planungs- und Forschungsvorhaben weiterverwenden und entsprechende Anfragen anderer

Organe der gesundheitssystemischen Selbstverwaltung realisieren. Dafür benötigen sie die Zustimmung ihrer jeweiligen Aussichtsbehörde. In diesem Rahmen sind komplexe Datenanalysen einschließlich Data-Mining möglich, sofern die Kassenärztlichen Vereinigungen ausreichend personelle Ressourcen bereitstellen können. In welchem Umfang sie ihre Datenbestände für komplexe Datenanalysen oder Data-Mining auf Anfrage anderer gesundheitssystemischer Organe oder aus Eigeninteresse tatsächlich weiterverwenden, könnten die jeweiligen Aufsichtsinstitutionen, nicht aber außenstehende Dritte beurteilen. Dem TAB ist nicht bekannt, welche finanziellen und personellen Ressourcen die KVen für analytische Datenweiterverwendungen bereitstellen können, da Auflistungen/Register zu Untersuchungsanfragen oder dadurch erzielte Ergebnis- oder Publikationslisten nicht veröffentlicht werden. Die Weiterverwendung der umfangreichen und aktuellen Datenbestände der KVen ist für Außenstehende daher wenig transparent.

5.4 Arzneimittelversorgung: Akteure, Daten und deren Verwendungsmöglichkeiten

Die Abgabe von Arzneimitteln an Patient/innen bzw. Kund/innen wird im ambulanten Bereich über Apotheken organisiert (Ausnahme freiverkäufliche Substanzen). Wie alle medizinischen Einrichtungen werden Apotheken als Wirtschaftsbetriebe geführt. Anders als bei Arztpraxen gilt die Führung einer Apotheke als Gewerbe (der grundsätzlich eine Gewinnerzielungsabsicht unterstellt wird). Analog zum Arztberuf (Kap. 4.1.1) ist auch der Apothekerberuf ein freier Beruf (Dienstleistung höherer Art im Interesse der Allgemeinheit), die Berufsausübung an die Mitgliedschaft in der spezifischen Berufskammer gebunden (die Grundsätze und Pflichten bei der Berufsausübung über Berufsordnungen definiert). Apotheker/innen unterliegen einerseits der Schweigepflicht, die auch die Daten zur Arzneimittelabgabe an einzelne Kund/innen zusätzlich zu den Vorgaben der DSGVO besonders schützt. Andererseits werden diese Schweige- und Datenschutzpflichten gegenüber gesetzlich krankenversicherten Kund/innen bei ärztlich verordneten und damit rezeptpflichtigen Arzneimitteln gesetzlich begrenzt (§ 300 SGB V). Apotheken müssen sämtliche Rezeptdaten an die jeweilige Krankenkasse übermitteln, wenn sie einen Teil der Abgabepreise den jeweiligen GKK direkt in Rechnung stellen. Auch im Apothekenbereich gibt es unterschiedliche Vereinigungen und Institutionen mit gesetzlich definierten Aufgaben, die in diesem Rahmen spezielle Datenbestände aufbauen.

5.4.1 Vereinigungen, Register, Informationssysteme

Apotheker/innen sind zur Mitgliedschaft in der jeweils zuständigen Landesapothekerkammer verpflichtet, die u. a. Register zu allen öffentlichen Apotheken[143] führen. Auszüge aus den Landesregistern (Name, Anschrift und Institutionenkennzeichen der Apotheke) sind zu einem bundeseinheitlichen Apothekenverzeichnis zusammenzuführen und dem Spitzenverband Bund der Krankenkassen unentgeltlich bereitzustellen und regelmäßig zu aktualisieren. Die Krankenkassen dürfen dieses Verzeichnis nur zur Erfüllung ihrer Aufgaben verwenden und nicht weitergeben (§ 293 Abs. 5 SGB V). Die Landesapothekenkammern sind auch für die Ausgabe elektronischer Apothekerausweise verantwortlich, die perspektivisch für bestimmte eHealth-Anwendungen relevant sind (Medikamentationspläne, eRezepte). Die Landes- und Bundesapothekenkammern sind berufspolitische Interessenvertretungen der Apotheker/innen.

Wirtschaftliche und (gesundheits)politische Belange werden eher über separate Apothekerverbände organisiert. Die Mitgliedschaft in einem Verband auf Landesebene bzw. auf Bundesebene (Deutscher Apothekerverband – DAV) ist freiwillig. Apothekenkammern und -verbände bilden gemeinsam die Bundesvereinigung Deutscher Apothekerverbände (ABDA), ein gemeinnütziger Verein, der u. a. das vollständige Register aller in Deutschland über Apotheken verfügbaren Arzneimittel führt.[144] Ein Kernelement der Registrierung ist die eindeutige Produktidentifikation. Dafür ist die Informationsstelle für Arzneispezialitäten (IFA GmbH) geschaffen worden. Sie vergibt auf Herstellerantrag gegen Vorlage definierter Arzneimitteldaten (u. a. Name, Darreichungsform, Packungsgröße, Inhaltsstoffe, Wirkstoffklassifikation der WHO [Anatomisch-Therapeutisch-Chemisches Klassifikationssystem – ATC-Code], definierte Tagesdosen, Indikationsgebiete, Arzneimittelstatus,[145] Preis sowie umfangreiche Angaben zum Hersteller) einerseits die national relevante Pharmazentralnummer (PZN) und andererseits die EU-weit gültige Pharmacy Product Number (PPN).[146] Die Daten- und Softwarespezialinstitution des ABDA (Avoxa – Me-

143 Apotheken, die Arzneimittel ambulant abgeben, werden als öffentliche Apotheken bezeichnet. Die stationäre Arzneimittelversorgung wird ohne Rezeptdokumentation mittel Krankenhausapotheken organisiert und in der Regel im Rahmen der Fallpauschalen vergütet (bei sehr teuren Arzneimitteltherapien gibt es spezielle Abrechnungsverfahren).

144 http://abdata.de/datenangebot/abdamed/ (10.11.2021)

145 Der Arzneimittelstatus ist eine nach dem Gefährdungspotenzial differenzierende vierstufige Arzneimittelgruppierung (freiverkäuflich, apothekenpflichtig, verschreibungspflichtig, Betäubungsmittel).

146 Die PZN wird als 8-stellige Zahl fortlaufend vergeben (codiert folglich keinerlei produkt- oder herstellerspezifische Angaben). Die PZN wird in die 18-stellige PPN integriert, die neben der fortlaufenden Nummer auch einige produktspezifische Informationen direkt codiert; Rechtsgrundlage: Richtlinie 2011/62/EU zur Änderung der Richtlinie 2001/83/EG zur Schaffung eines Gemeinschaftskodexes für Humanarzneimittel hinsichtlich der Verhinderung des Eindringens von gefälschten Arzneimitteln in die legale Lieferkette).

diengruppe Deutscher Apotheker GmbH) nutzt dieses Register u. a. zur kontinuierlichen Weiterentwicklung des verbandsspezifischen Arzneimittelinformationssystems ABDAMED. Es hat Zusatzmodule zur patientenindividuellen Arzneimittelrisikoprüfung, zur Berechnung von Abgabepreisen gegenüber den Krankenkassen sowie für tagesaktuelle Arzneimittelinformationen (von Neueinführungen bis zu Produktrückrufen). Register und Informationssystem werden mittels kostenpflichtiger Nutzungslizenz bereitgestellt.

Auch das Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM) führt zwei arzneimittelrelevante Register: Zum einen wird das nationale Arzneimittelinformationssystem PharmNet.bund.de seit Jahren auf- und ausgebaut. Für alle national verfügbaren Arzneimittel sollen vielfältige Informationen schrittweise integriert werden (u. a. Gebrauchs- und Fachinformationen, Risiko-Management-Pläne, Ergebnisse klinischer Prüfungen, Assessment Reports) Zum anderen führt das BfArM das Melderegister zu unerwünschten Arzneimittelwirkungen (UAW). Die Relevanz dieser Register für Data-Mining-Prozesse wird im Anwendungsbeispiel Pharmakovigilanz (Kap. 5.5.3) veranschaulicht. Für die Arbeitsprozesse bei der Abgabe von Arzneimitteln in Apotheken sind diese Register und Informationssysteme nicht unmittelbar relevant.

5.4.2 Apothekenrechenzentren

Die von den Apotheken genutzten Arzneimittelinformationssysteme haben in der Regel keine derart ausgebauten administrativen Arbeitsbereiche, dass die einzelnen Apotheken die Abrechnung mit den über 100 verschiedenen gesetzlichen Krankenkassen effizient realisieren können. Zumal die Abrechnung bei verschreibungspflichtigen Arzneimitteln an GKV-Versicherte wegen diverser patientenseitiger Zuzahlungs- und Befreiungsregelungen, krankenkassenseitiger Rabattverträge sowie hersteller- und apothekenseitiger Ab- und Zuschläge eine hohe Komplexität erreicht hat. Zumeist beauftragen Apotheken spezialisierte Apothekenrechenzentren (ARZ) mit der Leistungsabrechnung gegenüber gesetzlichen Krankenkassen. Apotheken sind grundsätzlich frei, ob, und wenn ja, welches ARZ sie mit ihren Abrechnungen beauftragen. ARZ können deutschlandweit agieren und sind nicht ausschließlich auf Apothekenabrechnungen beschränkt (einige übernehmen auch Abrechnungsaufgaben für andere ambulante medizinische Einrichtungen wie z. B. Physiotherapeut/innen). ARZ gelten inzwischen als Finanzdienstleistungsinstitute und stehen unter Aufsicht der Bundesanstalt für Finanzdienstleistungsaufsicht (BaFin).[147]

Bei einigen großen ARZ laufen seit Jahren erhebliche Datenmengen zur ambulanten Arzneimittelabgabe im ersten Gesundheitsmarkt zusammen. Der

147 Derzeit gibt es deutschlandweit weniger als 20 ARZ (überwiegend als GmbH geführt). Analyst/innen rechnen mit einer stärkeren Zentralisierung des Abrechnungsgeschehens, kleinere ARZ hätten es zunehmend schwerer, die BaFin-Auflagen zu erfüllen.

Ursprung dieser Daten sind die von den Ärzte/innen überwiegend maschinell erstellten Rezepte (bisher papierbasiert, nach derzeitigem Planungsstand sollen elektronische Rezepte 2022 eingeführt werden).[148] Rezepte sind hochgradig standardisierte Dokumente, die Daten zu den abzugebenden Wirkstoffen bzw. Arzneimitteln sowie direktidentifizierende Angaben sowohl zu rezeptausstellenden Ärzte/innen als auch zu Patient/innen und deren Krankenkasse enthalten. Die Rezeptdaten werden bei der Abgabe ergänzt: einerseits um die PZN des jeweiligen Arzneimittels und das Abgabedatum, andererseits um die zu erstattenden Beträge und das Institutionenkennzeichen der Apotheke (Abb. 4.4 rechts). Bisher müssen die vervollständigten papierbasierten Kassenrezepte wieder digitalisiert werden, entweder bereits in der Apotheke oder im jeweils beauftragten ARZ, das die Rezepte etwa aller 2 bis 4 Wochen abholt. Werden Einträge beim Scannen nicht erkannt, erfolgt eine manuelle Ergänzung oder Korrektur von Einlesefehlern. Die in den ARZ zusammenlaufenden Rezeptdatensätze bilden sowohl das Verschreibungsverhalten niedergelassener Vertragsärzt/innen, als auch die Arzneimittelabgabe an gesetzlich Versicherte auf Einzelfallebene ab. In der Summe wird dadurch ein wichtiger Teil des nationalen Arzneimittelmarktes detailliert und hochgranular (bis runter auf die Packungsebene) abgebildet – insgesamt ca. 55 % aller in öffentlichen Apotheken verkauften Arzneimittel (Forschungsgruppe PMV 2010, S. 22).[149]

Der primäre Verwendungszweck dieser Rezeptdatensätze ist die anteilige Kostenerstattung durch die jeweiligen GKK. Dafür werden Gesamtrechnungen einerseits pro Apotheke und andererseits pro GKK automatisiert erstellt. Den Kassen werden sowohl die Gesamtabrechnungen als auch die einzelnen gesetzlich definierten Rezeptdatensätze für die bei ihnen versicherten Personen übermittelt (Abb. 4.4 rechts). ARZ bieten ihren Apotheken zunehmend Onlinezugriffmöglichkeiten auf deren Rezeptdatensätze sowie auf Analysetools zum Monitoring und zur Optimierung betrieblicher Geschäftsprozesse an.

ARZ müssen für gesetzlich definierte Zwecke (u. a. Sicherung von wirtschaftlichen Verordnungsweisen, Beratung von Vertragsärzten zur Wirtschaftlichkeit bei Verschreibungen, Arzneimittelvereinbarungen und Abgabevolumen) ausgewählte Versorgungsdatensätze auf Anforderung weiterleiten an (§ 300 Abs. 2 SGB V):

Apothekenrechenzentren haben als privatwirtschaftliche Finanzdienstleister eine Sonderstellung im nationalen Gesundheitssystem. Unklar ist, ob sie als private Unternehmen dem Sozialdatenschutz unterliegen bzw. das Sozialgeheim-

148 www.bundesgesundheitsministerium.de/e-rezept.html (10.11.2021)

149 Im Datenbestand nicht enthalten sind Arzneimittelabgaben durch Privatrezepte und der Verkauf verschreibungsfreier Arzneimittel, an deren Kosten sich die GKV nicht beteiligt – 2. Gesundheitsmarkt, der zunehmend durch Onlineapotheken bedient wird (ausführlichere Darstellung z. B. Bundesregierung 2019b).

nis zu wahren haben (ARZ werden im § 35 SGB I nicht explizit genannt). Einerseits dürfen sie die ihnen übermittelten Daten nur für die im SGB V bestimmten Zwecke verarbeiten, soweit sie dazu von einer berechtigten Stelle beauftragt wurden. Andererseits wird ihnen explizit gestattet, anonymisierte Daten auch für andere Zwecke zu verarbeiten und zu nutzen (§ 300 Abs. 2 SGB V). Der Grad der Anonymisierung und die zulässigen Zwecke werden nicht näher benannt. Dadurch sind unterschiedliche Auslegungen möglich, die komplexe Datenanalysen zulassen oder begrenzen.

5.4.3 Das Deutsche Arzneiprüfinstitut

Das Deutsche Arzneiprüfungsinstitut (DAPI) ist ein von den Apothekenkammern und -verbänden gemeinsam getragener gemeinnütziger Verein zur Förderung von Wissenschaft und Forschung sowie zur Verbesserung der Arzneimitteltherapiesicherheit und Arzneimittelversorgung. Gegenwärtig sind sechs überregionale und vergleichsweise große ARZ Mitglieder des DAPI. Sie liefern seit der Jahrtausendwende von inzwischen mehr als 80 % der Apotheken Deutschlands versicherten-, arzt- und apothekenanonymisierte Datensätze auf Einzelrezeptebene (Rechtsgrundlage § 300 Abs. 2 SGB V).[150] Die Daten werden monatlich über die Datenannahmestelle (Avoxa – Mediengruppe Deutscher Apotheker GmbH) in das Data Warehouse des DAPI eingespeist, dort u. a. mit dem Arzneimittelregister des ABDA verknüpft und spätestens 8 Wochen nach Beendigung des Abrechnungsmonats für Analysen bereitgestellt (Schubert et al. 2014, S. 69). Jährlich kommen 600 Mio. Datensätze hinzu. In den ersten Jahren wurden die Datensätze mehrfach pseudonymisiert und zeitlich kontinuierlich fortgeschrieben. 2012 wurde das Verfahren aus datenschutzrechtlicher Perspektive neu bewertet und die zeitliche Fortschreibung unterbunden. Seit dem dürfen ARZ nur vollständig anonymisierte Datensätze an das DAPI übermitteln. Von den Versicherten werden lediglich Geburtsjahr und Versichertenstatus, von den Ärzte/innen die KV-Region (meist Bundesland) und das Verordnungsdatum übermittelt. Da zeitliche Fortschreibungen unmöglich sind, kann das DAPI seit 2012 keine Zeitverlaufs-/Längsschnittanalysen mehr durchführen (GKK können das mit ihren Rezeptdaten nach wie vor tun, die haben jedoch keine so große Marktabdeckung wie das DAPI; Kap. 5.5).

Datenanalysen werden nur von DAPI-Angehörigen durchgeführt, Ergebnisse nochmals auf Anonymität geprüft. Seinen Trägern und Mitgliedsorganisationen stellt das DAPI regelmäßig aktualisierte Basis- und Routineauswertungen bereit (u. a. Marktübersichten zur Abgabe von Impfstoffen, Substanz- oder Arzneimittelgruppen). Auch für Dritte (u. a. Institutionen des Gesundheitssys-

150 www.dapi.de/das-dapi/das-dapi-stellt-sich-vor/ (10.11.2021)

tems, Ministerien, Abgeordnete oder öffentliche Forschungseinrichtungen) führen sie Analysen auf Antrag durch (z. B. zur Entwicklung der Abgabe bestimmter Arzneimittelgruppen oder zu sich abzeichnenden Versorgungsengpässen), Kosten werden bisher nicht in Rechnung gestellt. Ergebnisse dürfen nur mit schriftlicher Genehmigung des DAPI an Dritte weitergegeben werden. Teilweise werden sie in Fachzeitschriften publiziert und über die DAPI-Website zugänglich gemacht.

5.4.4 Exkurs: kommerzielle Datenweiterverwendung – ein zulässiges Geschäftsmodell?

Vor einigen Jahren starteten zwei international tätige Wirtschaftsberatungsunternehmen mit Niederlassungen in Deutschland Initiativen zu Geschäftsanalysen im Arzneimittelmarkt (Kunze 2013; Machotta 2013). Sie schlossen Verträge zur Datenübermittlung sowohl mit Apothekenrechenzentren als auch mit einzelnen Apotheken und Arztpraxen ab, stellten ihnen einen schlüsselunabhängigen Pseudonymisierungsalgorithmus zur Verfügung (Kap. 3.3.3), mit dem diese regelmäßig die identifizierenden Merkmale ihrer einzelfallbezogenen Leistungsdatensätze pseudonymisierten. Diese Datensätze wurden über Clearingstellen und Trustcenter an die Beratungsunternehmen weitergeleitet, die mit dem periodenübergreifenden Pseudonym ihre Datenbestände regelmäßig fortschreiben und dadurch ebenfalls die ambulante Versorgungssituation auf Einzelfallebene für alle teilnehmenden ARZ, Apotheken und Praxen kontinuierlich abbilden konnten (was dem DAPI wegen der großen Marktabdeckung seit 2012 nicht mehr gestattet wurde, der KBV jedoch erlaubt ist). Die Wirtschaftsberatungsunternehmen beriefen sich auf faktische Anonymisierung der Daten und die Verfahrensgenehmigung durch die jeweils zuständige regionale Datenaufsichtsbehörde. ARZ ist die Weiterverwendung anonymisierter Daten für andere Zwecke explizit erlaubt (§ 300 Abs. 2 SGB V). Bei den Verträgen mit Apotheken und Arztpraxen wurde teils argumentiert, dass kein Kaufpreis pro Datensatz, sondern eine Aufwandsentschädigung für die Zusammenstellung und Anonymisierung der Daten gezahlt worden sei und die jeweiligen Einrichtungen spezielle, auf sie zugeschnittene Analysen und Auswertungen zu ihren Behandlungs- bzw. Verkaufsleistungen erhielten. Nähere Informationen zu den vertraglichen Vereinbarungen seien Geschäftsgeheimnisse.

Bei diesem Geschäftsmodell, bei dem sowohl ARZ als auch Apotheken und Arztpraxen faktisch anonymisierte Datensätze, die im Rahmen der medizinischen Versorgung von Patient/innen entstanden, mit Billigung der zuständigen Landesdatenschutzbeauftragten an private Marktforschungsinstitute veräußerten, wurde einerseits die unmittelbare Rechtmäßigkeit des Verfahrens kontrovers diskutiert. Andererseits wurde auch grundsätzlicher hinterfragt, wer Daten,

die im Kontext einer Behandlung entstanden sind, in welcher Form verarbeiten, nutzen oder auch verwerten darf.

Kritiker/innen dieses Vorgehens hielten die Daten nur für pseudonymisiert und wiesen darauf hin, dass Datensätze nie für sich alleine stünden und sich durch das Kombinieren von mehreren anonymisierten Datensätzen Personenbezüge wiederherstellen lassen (Reidentifizierungsrisiken; Kasten 3.7). Sie kritisierten auch, dass ursprünglich personenbezogene Daten besonderer Kategorie zu kommerziellen Transaktionen führten, ohne dass betroffene Patient/innen bzw. Kund/innen informiert und an den Erlösen beteiligt wurden. Der gesamten Transaktion würde es an Transparenz mangeln (Kunze 2013).

Grundsätzlich ist anzumerken, dass Verfahren zur Pseudonymisierung, Anonymisierung oder Löschung eine Veränderung oder Verarbeitung von Daten ist (Art. 4 DSGVO). Jegliche Verarbeitung personenbezogener Daten ist an einen Zweck gebunden und nur zulässig mit freiwilliger Einwilligung der betroffenen Person oder zur Wahrnehmung einer gesetzlich definierten Aufgabe im öffentlichen Interesse (z. B. Gesundheitsinteressen). Einrichtungen mit Aufgaben zur sozialen Sicherung wird die Befugnis zur Anonymisierung generell erteilt (§ 67c SGB X). Spezifische Regelungen konkretisieren, welche Aufgaben die Einrichtungen haben und für welche Zwecke sie die Daten nutzen dürfen. ARZ wird die Nutzung anonymisierter Daten für andere als im Sozialgesetzbuch bestimmte Zwecke explizit gewährt (§ 300 SGB V). Damit stellen sich einige Fragen: Warum wird dieses Recht so explizit nur ARZ gewährt? Auch könnte gefragt werden, ob Apotheken, die z. B. ihre Leistungen mit den GKK selbst abrechnen, die gleichen Rechte haben wie ARZ? Und wenn ja, ob das nur Apotheken betrifft oder alle medizinischen Einrichtungen als Leistungserbringer?

Grundsätzlich unterliegen sowohl Ärzte/innen als auch Apotheker/innen der Schweigepflicht. Aus medizinethischer/-rechtlicher Sicht wird argumentiert, dass die Schweigepflicht im medizinischen Bereich ein überragend wichtiges Gut schützt: Vertrauen (Krahnert 2016). Patient/innen sollen darauf vertrauen können, dass jegliche Informationen, die sie preisgeben und Daten, die im Rahmen der Behandlung erhoben werden, vertrauensvoll behandelt und geschützt werden. Ärzte/innen können ihre beruflichen Aufgaben (die Gesundheit ihrer Patient/innen zu erhalten und wiederherzustellen) nur erfüllen, wenn Patient/innen ihnen möglichst viele Daten und Informationen zu ihrer Person preisgeben. Deshalb sind die durch die Schweigepflicht geschützten medizinischen Behandlungsdaten nicht nur durch das allgemeine Persönlichkeitsrecht geschützt, sondern bleiben auch staatlichem Zugriff weitgehend verschlossen (nicht einmal im Rahmen staatlicher Überwachungsmaßnahmen der Strafprozessordnung darf auf diese Daten zugegriffen werden). Jedoch durchbrechen inzwischen zahlreiche Einzelgesetze die Schweigepflicht und definieren die

Weitergabe patientenbezogener Daten, um sowohl medizinische als auch wirtschaftliche Prozesse organisieren zu können. Zudem unterliegen anonymisierte Daten keinen Schweige-/Datenschutzpflichten mehr, betroffene Personen haben keine Rechte an anonymisierten Daten (Metschke/Wellbrock 2002, S. 20).

Analysen zu Wirtschaftlichkeitsprozessen liegen im Grenzbereich des jeweiligen Berufsethos. Einerseits soll das berufliche Handeln von Ärzte/innen und Apotheker/innen am Patientenwohl und nicht am Gewinnstreben ausgerichtet sein. Dem Vertrauen, das den Berufsständen entgegengebracht wird, solle entsprochen werden (z. B. Bayrische Landesapothekerkammer 2017; M-BOÄ). Andererseits ist der Betrieb einer Apotheke ein Gewerbe, also eine wirtschaftliche Tätigkeit mit Gewinnerzielungsabsicht.

Nach einer auf dem Deutschen Ärztetag 2014 geführten Debatte sprach sich die Mehrheit der teilnehmenden Ärzte/innen dafür aus, auch anonymisierte Behandlungsdaten nicht ohne Patienteneinverständnis für andere Zwecke zu nutzen oder gar zu verkaufen. Da mit den zunehmenden Möglichkeiten der Reidentifizierung die Gefahr des Missbrauchs steigt, würde dies das Vertrauensverhältnis zwischen Ärztin/Arzt und Patient/in grundsätzlich gefährden. Ein diesbezüglicher Entschließungsantrag wurde auf dem Ärztetag angenommen (Bodammer et al. 2014, S. 237). In der Apothekerschaft wurde bisher keine vergleichbare Diskussion bekannt.

5.4.5 Einschätzung

Trotz des noch bestehenden Medienbruchs wegen der papierbasierten Rezeptausstellung haben die abrechnungsrelevanten Rezeptdatensätze einen hohen Grad an semantischer und syntaktischer Interoperabilität. Alle Angaben sind codiert, über Arzneimittelregister können vielfältige Zusatzangaben zu einzelnen Substanzen ergänzt werden. Zudem haben sie wegen der hohen Auflösung sowie der Raum-, Zeit- und Personenbezüge ein erhebliches datenanalytisches Potenzial (Anwendungsbeispiel in Kap. 5.5.3).

Apotheken und deren Rechenzentren haben im öffentlichen Gesundheitssystem eine gewerbliche Sonderstellung, die auch im Umgang mit den Register- und Rezeptdaten sichtbar wird (Verfahren sind weitgehend kommerzialisiert). Apothekenrechenzentren haben sich im Abrechnungsprozess ambulanter Arzneimittelverschreibungen zu zentralen Datendrehscheiben entwickelt (ähnlich wie Kassenärztliche Vereinigungen bei der Abrechnung ambulanter Behandlungsleistungen). Während das SGB V die Möglichkeiten und Grenzen von Wirtschaftlichkeitsprüfungen für Kassenärztliche Vereinigungen dezidiert vorgibt, bleiben diesbezügliche Vorgaben für Apothekenrechenzentren vergleichsweise vage (§ 300 Abs. 2 SGB V). Als expliziter Finanzdienstleister bieten die Rechenzentren über ihre Plattformen ihren Kunden auch Analysetools zur Optimierung betrieblicher Geschäftsprozesse und ergänzen damit die eher schwach

ausgebildeten administrativen Bereiche der Arzneimittelinformationssysteme von Apotheken. Es gibt kaum aktuelle Hinweise, inwiefern sich externe Wirtschaftsberatungsunternehmen mit solchen Analysetools in nennenswertem Umfang ebenfalls dauerhaft am Markt platzieren konnten.

Beim Umgang mit faktisch anonymisierten Daten wird der Interpretationsspielraum der rechtlichen Vorgaben sowohl im Selbstverständnis der datenverarbeitenden Stellen als auch bei der Datenschutzaufsicht sichtbar. Es werden datenanalytische Möglichkeiten in unterschiedlichem Maße begrenzt. Während sich Ärzte/innen grundsätzlich dagegen aussprachen, anonymisierte Patientendaten kommerziell weiterzuverwenden, werden im Arzneimittelbereich diese Datenverwertungsmöglichkeiten genutzt. Auch bezüglich der Frage wer, welche faktisch anonymisierten Datenbestände fortschreiben und weiterverwenden darf, gab es in der Vergangenheit unterschiedliche Einschätzungen seitens der Datenaufsicht.

5.5 Gesetzliche Krankenkassen: Daten und Analysemöglichkeiten

5.5.1 Aufgaben, Strukturen, Datenbestände

Die gesetzlichen Krankenkassen bzw. Ersatzkrankenkassen (vereinfachend gemeinsam als GKK bezeichnet) versichern als Träger der gesetzlichen Krankenversicherung (GKV) und der gesetzlichen Pflegeversicherung Personen gegen gesundheitliche Risiken (*Versicherungsverhältnis*). Sie sollen die Gesundheit der Versicherten erhalten und soweit möglich wiederherstellen. Dazu tragen sie zum einen die Kosten für notwendige medizinische und pflegerische Leistungen, die Ärzte/innen und andere Fachkräfte in medizinischen Einrichtungen in ihrem Auftrag erbringen (*Geschäftsverhältnis*). Dieses Dreiecksverhältnis der medizinischen Versorgung wird auch als erster Gesundheitsmarkt bezeichnet. GKK sollen zudem die gesundheitliche Eigenverantwortung und -kompetenz ihrer Versicherten fördern (Prävention). Diesbezüglich können sie selbst Leistungen anbieten oder Dritte (in der Regel keine medizinischen Einrichtungen) damit beauftragen. Das Versicherungsverhältnis wird bestimmt durch

- das *Solidaritätsprinzip,* d. h., der Beitrag jedes Versicherten bemisst sich nach dem persönlichen Einkommen und nicht nach dem Krankheitsrisiko und
- ein weitgehendes *Sachleistungsprinzip*, d. h., Versicherte erhalten medizinische und andere Leistungen (kein Geld). Ein Teil der Leistungen ist gesetzlich definiert (auch als Pflichtleistungen oder Regelversorgung bezeichnet), wobei die GKK hier teilweise einen Ermessensspielraum haben (u. a. Leistungserstattung auf Antrag; Kap. 4.3.3). Bei Aufgaben zur Förderung

der gesundheitsbewussten Lebensweise können GKK ihr Leistungsangebot selbst definieren (Satzungsleistungen).

Das Geschäftsverhältnis zu den medizinischen Einrichtungen als Leistungserbringern wird durch das SGB V sowie konkretisierende Richtlinien und Verträge definiert. Grundsätzlich dürfen GKK nur Leistungen erstatten, die notwendig, ausreichend, zweckmäßig und wirtschaftlich sind (§ 12 SGB V). Welche Leistungen das sind, wird wesentlich durch den Gemeinsamen Bundesausschuss (G-BA) und dessen Richtlinien entschieden.[151] Im Rahmen des Geschäftsverhältnisses prüfen GKK u. a. die Regelkonformität der in Rechnung gestellten medizinischen Leistungen (bei ambulant erbrachten medizinischen Leistungen zusammen mit den Kassenärztlichen Vereinigungen).

Im Rahmen der Versicherungs- und Geschäftsbeziehungen verwalten die GKK unterschiedliche Datenbestände. Zudem sind sie an unterschiedlichen gesundheitssystemischen Aufgaben beteiligt, für die sie Datensätze liefern und/oder Analysen durchführen.

Nummern, Datenbestände, Nutzungsmöglichkeiten

Zum einen hat jede GKK ein eindeutiges Institutionskennzeichen (IK), das im Rahmen der Selbstverwaltung u. a. für die eindeutige Identifizierung einzelner Kassen und die Pseudonymisierung GKK-bezogener Daten genutzt wird. Zum anderen vergeben die GKK an ihre Versicherten eine eindeutig identifizierende *Krankenversicherungsnummer* (KV-Nr.) und eine elektronische Gesundheitskarte (eGK), die derzeit im Wesentlichen eine Ausweisfunktion im Rahmen der GKV hat. Zudem führt jede GKK ihr Versichertenverzeichnis mit den direkt identifizierenden (Stamm-)Daten (Name, Anschrift, Alter, Geschlecht, KV-Nr.), den Daten zum Versichertenstatus (u. a. Zuzahlungsstatus, Wahltarife) und den Angaben zur Beitragsbemessung enthält. Diese Verzeichnisse müssen die Krankenkassen besonders schützen von anderen Datenbeständen getrennt aufbewahren. Sie werden weder zentral zusammengeführt noch an Dritte weitergegeben.[152]

151 Der Gemeinsame Bundesausschuss (G-BA) ist das oberste Entscheidungsgremium der gesundheitssystemischen Selbstverwaltung (§ 91 SGB V). Seine 13 stimmberechtigten Mitglieder setzen sich aus GKK-Vertreter/innen als Leistungsträger (GKV-Spitzenverband) und Vertretungen der medizinischen Leistungserbringer (KVen/KBV, Deutsche Krankenhausgesellschaft) zusammen. Patientenvertretungen haben kein Stimmrecht.

152 Jede GKK richtete 1995 eine eigenständige rechtsfähige Pflegekasse ein, um Leistungen im Rahmen der gesetzlichen Pflegeversicherung separat abwickeln zu können (Rechtsgrundlage: SGB XI). Kranken- und Pflegekassen können Versichertenverzeichnisse im Rahmen ihrer gesetzlichen Aufgaben gemeinsam nutzen (§§ 46 und 96 SGB XI), Krankenkassen und Unfall- oder Rentenversicherungsträger dürfen das nicht. Letztere vergeben eigenständige Identifikationsnummern und führen eigene Register mit den Stammdaten ihrer Versicherten.

(Fach-)Datenbestände sind vor allem die versicherten- bzw. fallbezogenen Leistungsabrechnungsdaten der medizinischen Einrichtungen (Krankenhäuser, ambulante Praxen, Apotheken), die ihre Versicherten behandelt haben. Einen weiteren Datenbestand bilden die versichertenbezogenen Daten im Kontext von Satzungsleistungen (z. B. zur Realisierung von Bonusprogrammen). Jeder Datenbereich muss getrennt gespeichert und verarbeitet werden. Direkt- und quasiidentifizierende Merkmalsbereiche sind mit der KV-Nr. zu pseudonymisieren.

GKK sind Sozialgeheimnisträger, die u. a. ihrer Aufsichtsbehörde eine jährliche Übersicht über die Art der gespeicherten Sozialdaten vorlegen müssen (Kap. 5.1). Sie sind einerseits zur Datensparsamkeit, ausschließlich zweckgebunden Nutzung und Anonymisierung/Löschung nach spätestens nach 4 Jahren verpflichtet, wobei Daten, die relevant für den Erhalt späterer Leistungen sein können, bis zu 10 Jahren aufbewahrt werden dürfen (§ 292 SGB V). Andererseits dürfen GKK trotzdem zunehmend große Datenbestände aufbauen und sind Verarbeitungszwecke z. T. weit definiert (§ 284 SGB V):

- Verwaltungsaufgaben im Rahmen des Versicherungsverhältnisses (u. a. Feststellung des Versicherungsverhältnisses, Ausstellung der elektronischen Gesundheitskarte);
- Leistungsabrechnung, einschließlich Plausibilitäts- und Rechtmäßigkeitsprüfungen;
- Überwachung der Wirtschaftlichkeit (nur Stichproben in notwendigem Maße);
- Planung und Durchführung von Modellvorhaben der medizinischen Versorgung;
- Unterstützung der Versicherten bei Behandlungsfehlern sowie
- andere durch Rechtsvorschriften des Sozialgesetzbuches angeordnete oder erlaubte Zwecke (z. B. Bonusprogramme entsprechend der jeweiligen Satzung). Diese Klausel, legitimiert über nachgeordnete, externe Rechtsvorschriften weitere datenanalytische Möglichkeiten für GKK im Rahmen öffentlicher Aufgaben und kann komplexe Rechtsstrukturen der Datenweiterverwendung nach sich ziehen, die für betroffene Personen schwer nachvollziehbar sein dürften.

Zudem dürfen GKK ihre Daten über die normalen Fristen hinaus aufbewahren und für zeitlich befristete und im Umfang begrenzte Forschungsvorhaben mit Erlaubnis der Aufsichtsbehörde leistungserbringer- oder fallbeziehbar selbst auswerten (insbesondere um epidemiologische Erkenntnisse, Informationen über örtliche Krankheitsschwerpunkte oder über Zusammenhänge zwischen Erkrankungen und Arbeitsbedingungen zu gewinnen). Dafür wird zwar die Anonymisierung der Datenbestände gefordert (§ 287 SGB V, ähnlich § 67c SGB X), da jedoch einzelfallbeziehbare Analysen möglich bleiben sollen, ist höchstens

eine schwache Anonymisierung möglich, die bei längeren zeitlichen Fortschreibungen teilweise infrage gestellt wird (Kasten 3.7). Die Aufsichtsbehörden der GKK agieren in dieser Struktur als externe Data Access Committees, die die Einhaltung ethischer Forschungsstandards bei der Weiterverwendung von Sozialdaten zu Forschungszwecken im Einzelfall prüfen und sichern sollen. Einwilligungsmanagementsysteme, die vor der formalen Anonymisierung ansetzen müssten, sind dem TAB nicht bekannt.

Organisation und Finanzierungsverfahren

In Deutschland sind ca. 90 % der Bevölkerung (73,3 Mio. Personen) bei einer von derzeit gut 100 GKK versichert.[153] GKK agieren als Körperschaften öffentlichen Rechts organisatorisch und finanziell selbstständig in historisch gewachsenen Strukturen (Schepers et al. 2015, S. 148 f.): Die Allgemeinen Ortskrankenkassen (AOK) haben eine hierarchische Struktur mit 11 regional eigenständigen AOK-Unternehmen und einem AOK-Bundesverband (insgesamt 24 Mio. Versicherte), zu dem u. a. ein wissenschaftliches Institut der Ortskrankenkassen (WIdO) gehört. Die Deutsche Rentenversicherung Knappschaft-Bahn-See ist sowohl eine GKK (1,6 Mio. Versicherte) als auch eine Berufsgenossenschaft (Unfallversicherung), Trägerin der gesetzlichen Rentenversicherung (1,7 Mio. Personen) und diverser medizinischer Einrichtungen (11 Krankenhäuser, 11 Rehakliniken, 1.500 Knappschaftsärzte mit eigenem sozialmedizinischem Dienst). Neben diesen zwei besonderen Organisationsformen gibt es acht weitere große GKK (jeweils 1 bis 10 Mio. Versicherte deutschlandweit), die ähnlich der AOK spezifische Datenanalyseabteilungen eingerichtet haben. Dazu kommt eine Vielzahl kleiner, ursprünglich auf einzelne Firmen oder Branchen begrenzte Krankenkassen (teilweise nur mit wenigen Tausend Versicherten). Insbesondere bei diesen kleineren Kassen gab es in den vergangenen Jahren vielfach Zusammenschlüsse, aber auch Insolvenzen.

GKK haben Verbände auf Landes- und Bundesebene, um Interessen zu bündeln und gemeinsam vertreten zu können sowie Aktivitäten abzustimmen. Der 2007 eingerichtete Spitzenverband Bund der Krankenkassen ist der zentrale Verhandlungspartner auf Bundesebene (§ 217 SGB V), der u. a. den Qualitäts- und Wirtschaftlichkeitswettbewerb der GKK untereinander organisieren und die GKK in Bezug auf den elektronischen Datenaustausch (Interoperabilität und Vernetzung) (§ 217f Abs. 2 SGB V) unterstützen soll und sich zunehmend zu einer zentralen Datendrehscheibe des Gesundheitssystems entwickelt.

153 10 % der Bevölkerung (ca. 9 Mio. Einwohner) sind derzeit privat krankenversichert (Beamte, Freiberufler, teilweise Selbstständige sowie Personen und Angestellte hoher Vergütungsstufen). Private Krankenversicherungen (PKV) werden von 43 Unternehmen angeboten. Bei der PKV gilt das Äquivalenzprinzip, d. h., persönliche Krankheitsrisikofaktoren (Gesundheitsstatus, Alter) beeinflussen die Beitragshöhe.

Regional tätige GKK werden von länderspezifischen Aufsichtsbehörden überwacht, überregional tätige vom Bundesamt für Soziale Sicherung (BAS), GKK-Verbände vom BAS oder dem BMG. Diese Aufsichtsbehörden prüfen u. a. die Rechtskonformität der Satzungen sowie die Betriebsführung der GKK und deren Verbände (§ 67c Abs. 2 SGB V).

GKK sind zur Kostendeckung verpflichtet. Gesetzlich festgelegte Versichertenbeiträge und ein Bundeszuschuss (für versicherungsfremde Leistungen wie z. B. die Mitversicherung von Familienangehörigen) bilden die wesentlichen Einkünfte jeder GKK. Können sie ihre Kosten damit nicht decken, müssen sie Zusatzbeiträge bei ihren Versicherten erheben. Personen, die von der GKV nicht de facto ausgeschlossen sind (u. a. Beamte), haben bezüglich der über 100 GKK ein Wahlrecht, GKK eine Aufnahmepflicht. GKK müssen als GKV-Träger medizinisch notwendige Leistungen finanzieren, wirtschaftlich haushalten und stehen aufgrund der Wahlfreiheit der Versicherten in einem gewissen Wettbewerb untereinander (Schepers et al. 2015, S. 177 f.). Kassen, deren Mitglieder im Durchschnitt weniger verdienen, einen höheren Krankenstand aufweisen oder kostenintensivere Erkrankungen erleiden, haben im Prinzip einen Wettbewerbsnachteil. Zusätzliche Angebote wie z. B. Bonusprogramme (Satzungsleistungen), die vorrangig gesunde Personen mit höheren Einkommen ansprechen, gelten als Wettbewerbsverstärker und werden seit Jahren kontrovers diskutiert.

Um den Wettbewerb zwischen den GKK um die gesündesten Versicherten zumindest abzumildern, wurde 1994 zunächst ein einfacher Finanzausgleich und 2009 der morbiditätsorientierte Risikostrukturausgleich (Morbi-RSA) und ein zentraler, beim BAS angesiedelter Gesundheitsfonds eingeführt. Seitdem fließen die Beiträge aller gesetzlich krankenversicherten Personen und der Bundeszuschuss zunächst in diesen Fonds[154] und werden dann anhand des jährlich für jede GKK neu zu berechnenden Risikostrukturausgleichs aufgeteilt. Mit diesem Verfahren soll die aufgrund der Versichertenstruktur bestehenden Finanzierungsrisiken jeder GKK ermittelt und ausgeglichen werden (Schepers et al. 2015, S. 179 ff.). Bei der Einführung dieserart Kassenrückversicherung suchte man Faktoren, die von den Kassen möglichst nicht beeinflusst und automatisiert erfasst werden konnten. Bei der Einführung entschied man, die im Rahmen der Leistungsabrechnung ohnehin anzugebenen Haupt- und Nebendiagnosen sowie ergänzend Arzneimittelverschreibungen zu nutzen und anhand dieser Daten 50 bis 80 Krankheiten zu definieren, die als Morbiditätskriterien in die Berechnung des Risikostrukturausgleichs einflossen. Die Festlegung der Krankheiten erfolgte nicht datenbasiert (nicht nur die kostenintensivsten, auch die häufigsten

154 Gesamtvolumen 2020: 265 Mrd. Euro: www.bundesgesundheitsministerium.de/fileadmin/Dateien/3_Downloads/G/GKV/210302_PM_Anlage_barrierefrei_1.-4._Qu._2020_bf_Tabelle.pdf (10.11.2021)

Krankheiten sollten berücksichtigt und für die Berechnung des Strukturausgleichs spezifisch gewichtet werden). Für die Berechnung Morbi-RSA erstellt jede GKK für das zurückliegende Abrechnungsjahr aus den unterschiedlichen Leistungsabrechnungsdaten einen Jahresdatensatz mit den definierten Daten für jeden Versicherten und übermittelt diesen an den Spitzenverband Bund der Krankenkassen, der diese Daten zusammenführt, um sie zum einen an das BAS zur Aktualisierung des Morbi-RSA weiterzuleiten und sie zum anderen im Rahmen der eigenen Aufgaben zu verwenden.

Auch dieses Verfahren wird wegen der jährlichen Fortschreibung als lernendes System bezeichnet, das seit der Einführung kontrovers diskutiert wird. Die angebliche Unbeeinflussbarkeit des Verfahrens wurde stets angezweifelt. Denn Krankenkassen begannen auf unterschiedliche Art und Weise medizinische Einrichtungen dahingehend zu beraten, dass sie die Haupt- und Nebendiagnosen im Rahmen ihres Entscheidungsspielraums möglichst morbi-RSA günstig codieren. Auch die mathematische Form der Gewichtungsfunktionen und die Komplexität des Verfahrens wurde kontrovers diskutiert (ausführlich z. B. Baas/Möws 2017). Gutachten im Rahmen der Begleitforschung wurden erstellt (Dietzel et al. 2015; Drösler et al. 2017) und mündeten in eine grundlegende Überarbeitung der Ausgleichsverfahren in 2020. Um sie resistenter gegen Manipulationen zu machen, sollen keine Krankheiten im Vorfeld definiert, sondern das gesamte Krankheitsspektrum berücksichtigt werden. Die Klassifikations- und Berechnungsmodelle werden um einiges komplexer, benötigen zusätzliche Daten und sollen perspektivisch engmaschiger evaluiert werden.[155]

5.5.2 Daten aus der Leistungsabrechnung: Bestandteile, Haltung, Mehrfachnutzung

Bis 2003 hatten GKK nur einen begrenzten Einblick in Daten, die zum einen den Gesundheitszustand ihrer Versicherten und zum anderen innerbetriebliche Prozesse medizinischer Einrichtungen abbilden (Schepers et al. 2015, S. 150). Durch das GKV-Modernisierungsgesetz[156] erhalten sie seit 2004 mit den standardisierten Datensätzen aus der Leistungsabrechnung einen tieferen Einblick sowohl in den Gesundheitszustand ihrer Versicherten als auch in Behandlungsprozesse einzelner Einrichtungen (Abb. 4.4):

Anhand der stationären Leistungsabrechnungsdaten erhalten GKK Informationen zur behandelnden medizinischen Einrichtung, zum Aufnahme-, Verlegungs- und Entlassungszeitpunkt, zum gesundheitlichen Zustand des Versicherten (anhand der gestellten Haupt- und Nebendiagnosen), zu medizinischen

155 www.bundesamtsozialesicherung.de/de/themen/risikostrukturausgleich/weiterentwicklung/ (10.11.2021)

156 Gesetz zur Modernisierung der gesetzlichen Krankenversicherung (GKV-Modernisierungsgesetz – GMG

Behandlungsmaßnahmen sowie zur Erstattungspauschale (§ 301 SGB V). Kaum enthalten sind Angaben zur Verabreichung von Arzneimitteln.[157] Nicht enthalten sind Angaben zur pflegerischen und sozialen Betreuung und Unterstützung oder zur Patientenzufriedenheit. Da Krankenhäuser erbrachte Leistungen direkt mit der jeweiligen GKK des Patienten abrechnen, müssen letztere die Leistungsdatensätze zunächst selbst auf Vollständigkeit und Richtigkeit prüfen und ggf. den Medizinischen Dienst hinzuziehen (Kap. 4.4.1), bevor sie anerkannte Leistungen vergüten und die Datensätze archivieren.

Auch wenn GKK im ambulanten Bereich die Abrechnung und Vergütung ärztlicher Leistungen weitgehend an Kassenärztliche Vereinigungen delegiert haben (Kap. 5.3), bekommen sie seit 2004 im Anschluss an den Leistungsabrechnungsprozesses von den KVen arzt- und versichertengenaue Leistungsdatensätze mit Haupt- und Nebendiagnosen, erbrachten ärztlichen Leistungen, Zeitangaben sowie den abgerechneten Gebührenpositionen (§ 295 SGB V). Auch diese Datensätze enthalten keine Angaben zu verordneten Arzneimitteln. Eine Sonderstellung nehmen besondere ambulante Versorgungsformen ein, die über Selektivverträge zwischen einzelnen GKK und medizinischen Einrichtungen vereinbart und über Jahre getestet werden (u. a. Disease Management Programme, Modellvorhaben der integrierten oder hausarztzentrierten Versorgung). Spezifisch definierte, umfangreichere Abrechnungsdatensätze werden direkt bei den jeweiligen GKK eingereicht, geprüft und Leistungen direkt vergütet, ohne dass Kassenärztliche Vereinigungen daran beteiligt sind.

GKK tragen zumindest einen Teil der Kosten verordneter rezeptpflichtiger Arzneimittel und erhalten für die Abrechnung standardisierte Rezeptdatensätze auf Einzelfallebene (Kap. 5.4). Der Einsatz von Heil- und Hilfsmitteln sowie digitaler Gesundheitsanwendungen wird ähnlich abgerechnet (§§ 301a u. 302 SGB V), sodass GKK auch diesbezüglich versicherten- und einrichtungsgenaue Leistungs(abrechnungs)daten haben.

Parallel dazu laufen bei den GKK auch die Leistungs(abrechnungs)daten der gesetzlichen Pflegeversicherung (§ 28 SGB X]) zusammen.

Zudem bekommen GKK zur internen aufgabenbezogenen Verwendung diverse Register, Verzeichnisse und Klassifikationen (u. a. IK-Register, Arzt-, Betriebsstätten- und PZN-Verzeichnisse, medizinische Klassifikation und Gebührenkataloge), anhand derer sie ihre codierten und teilweise pseudonymisierten Daten aus der Leistungsabrechnung mit weiteren Datensätzen relational verknüpfen können.

157 Im Normalfall sind die Kosten des Arzneimitteleinsatzes in der DRG-Fallpauschale inkludiert. Nur der Einsatz sehr teurer Präparate wird ergänzend zu den Pauschalen über Zusatzentgelte erstattet und dadurch separat ausgewiesen (InEK 2021, S. 132 ff.; Schepers et al. 2015, S. 152).

Diese hochgradig standardisierten Daten aus der Abrechnung unterschiedlicher medizinischer Versorgungsleistungen werden teilweise auch als *Routinedaten* bezeichnet. Grundsätzlich verwaltet jede GKK nur die Daten ihres Versichertenkollektivs. Bei großen überregionalen GKK kommen jedoch regelmäßig Daten von vielen medizinischen Einrichtungen und vielen Versicherten zusammen. Diese Situation wurde u. a. bei der Verabschiedung des GKV-Modernisierungsgesetzes kontrovers diskutiert, zumal diese Daten zunehmend auf medizinischen Konzepten aufbauen und erhebliche Einblicke in den Gesundheitszustand der Versicherten ermöglichen. Auch können Versorgungsprozesse medizinischer Einrichtungen genauer abgebildet und verglichen werden. Im Zentrum der Debatte stand der/die »gläserne Patient/in bzw. Versicherte« und die ungenügende Pseudonymisierung der versichertenbezogenen Daten. Diese Auseinandersetzung führte zum einen zum Verbot der sektorübergreifenden Datenzusammenführung, was GKK durch technische und organisatorische Maßnahmen sicherstellen müssen.[158] Bis heute dürfen GKK keine durchgehende Versichertenakte führen. Sie speichern Leistungsabrechnungsdaten von Krankenhäusern, Arztpraxen, Apotheken und weiteren medizinischen Einrichtungen des ambulanten Bereichs grundsätzlich getrennt. Zum zweiten wurde eine strikte Zweckbindung für die Datennutzung im Rahmen der primären GKK-Aufgaben festgeschrieben (u. a. werden Form und Umfang von Wirtschaftlichkeitsuntersuchungen dezidiert definiert [§§ 106 ff. SGB V]).

Jenseits ihrer primären Aufgaben können GKK die Datenverwendung für begrenzte Forschungsvorhaben bei ihrer Aufsichtsbehörde beantragen. Dadurch wurden in der Vergangenheit bereits unterschiedliche komplexe datenanalytische Projekte unter Verwendung von Routinedaten von einzelnen großen GKK realisiert. Beispiele sind:

- *Arzneimittelverordnungs-Report*: Ursprung ist ein in den 1980er Jahren von der AOK, dem WIdO und Wissenschaftler/innen akademischer Einrichtungen initiiertes und vom BMBF gefördertes Projekt, durch das Verfahren entwickelt wurden, mit denen die bei der AOK zusammenlaufenden Rezeptdaten regelmäßig analysiert werden. Primäres Ziel dieser Analysen ist die Verbesserung der Markt- und Kostentransparenz. Im Zentrum steht das Verordnungsverhalten bei neuen Arzneimitteln. Die Analyseverfahren werden bis heute genutzt und weiterentwickelt, Ergebnisse und abgeleitete Erkenntnisse jährlich publiziert.[159] Kritisch werden mitunter aus den Ergebnissen abgeleitete Verordnungsempfehlungen gesehen, weil Krankenkassen die Therapiefreiheit von Ärzt/innen begrenzen würden.

158 www.bfdi.bund.de/DE/Buerger/Inhalte/GesundheitSoziales/Allgemein/DatenuebermittlungZuAbrechnungszwecken.html (10.11.2021)

159 www.wido.de/publikationen-produkte/buchreihen/arzneiverordnungs-report/ (9.11.2021)

- *Analysen zur Qualitätssicherung* (QS): Ziel des 2002 vom AOK-Bundesverband initiierten Projekts war die Entwicklung von Verfahren zur Messung der Qualität medizinischer Versorgung anhand von Routinedaten, wobei auch im späteren Therapieverlauf auftretende Ereignisse berücksichtigt werden sollten. Die Verfahren nutzt die AOK, Ergebnisse speist sie in ihren AOK-Krankenhausnavigator ein.[160] In die nationalen QS-Strukturen wurden diese vergleichsweise aufwandarmen Verfahren bisher nicht integriert (Kap. 4.4.2).
- *Polypharmazieprojekt »AdAM«* (Anwendung digital-unterstütztes Arzneimitteltherapie- und Versorgungsmanagement): Ziel des seit 2017 durch den Innovationsfonds der GKV geförderten Projektes ist die Entwicklung einer digitalen Anwendung, die Ärzt/innen, Apotheker/innen und Versicherte auf mögliche Wechselwirkungen und Risiken von Arzneimittelkombinationen hinweist.[161] Dazu werden aus den Rezeptdaten einer großen, landesweit agierenden GKK Kombinationen von gleichzeitig verordneten Arzneimitteln herausgefiltert und eine Datenbank aufgebaut, in der diese Kombinationen mit Zusatzinformationen zu Risikopotenzialen angereichert werden. Bei neuen Kombinationsverordnungen wird gezielt nach den Arzneimittelkombinationen gesucht und ggf. Zusatzinformationen sowie Risiko-/Warnhinweise ausgegeben (ähnlich werden z. B. die Tumormutationsregister aufgebaut und genutzt; Kap. 4.3.3). Seit Jahren gibt es vielfältiges Engagement zum Auf- und Ausbau von Datenbanken mit Arzneimittelwechselwirkungen und der Weiterentwicklung von Arzneimittelinformationssystemen (Kap. 5.4.1).

Datenzusammenführung, Datentransparenz und Forschungsdatenzentrum

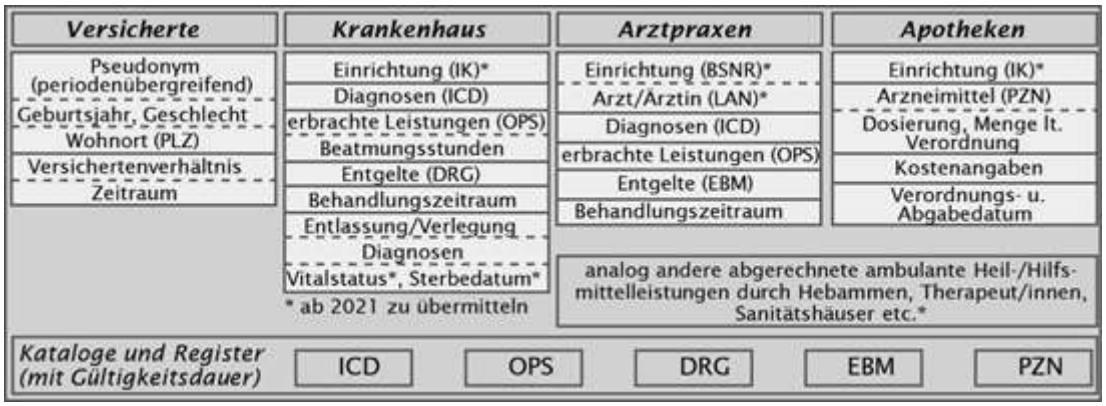
Nach der Jahrtausendwende wurde der Begriff der *Datentransparenz* im SGB V verankert. Dort steht er für die deutschlandweite Zusammenführung aller Leistungs(abrechnungs)datensätze auf Einzelfallebene von allen GKK und für die Nutzung dieser Daten zur Planung, Steuerung und Weiterentwicklung gesundheitssystemischer Teilbereiche sowie für Sozial- und Versorgungsforschungszwecke (Schepers et al. 2015, S. 179 ff.). Es dauerte zunächst Jahre, bis man sich auf die dazu erforderlichen Verfahren verständigt hatte. Die erste Datentransparenzverordnung (DaTraV)[162] wurde neun Jahre später vorgelegt. Der Datensatz

160 www.qualitaetssicherung-mit-routinedaten.de/ (10.11.2021)
161 innovationsfonds.g-ba.de/projekte/neue-versorgungsformen/adam-anwendung-digital-gestuetztes-arzneimitteltherapie-und-versorgungs-management.71 (10.11.2021).
162 Verordnung zur Umsetzung der Vorschriften über die Datentransparenz (Datentransparenzverordnung – DaTraV)

zur Fortschreibung des Morbi-RSA wurde als maßgeblicher Datenpool bestimmt und inhaltlich erweitert (Abb. 5.2). Die Datenzusammenführung verläuft in mehreren Schritten: Jede GKK stellt bis zum Ende des 3. Quartals einen versicherten-, Ärzte/innen-und einrichtungspseudonymisierten Jahresdatensatz mit allen Leistungsabrechnungsdaten auf Einzelfallebene des Vorjahres zusammen (§ 303b SGB V i. V. m. § 3 DaTraV) und übermittelt ihn an den Spitzenverband Bund der Krankenkassen (Datensammelstelle), der diese Datensätze ursprünglich sofort an das Bundesamt für Soziale Sicherung (BAS) weiterleitete. Das BAS prüfte und korrigierte die Daten in einem mehrstufigen Verfahren und fügte sie zu einem Gesamtjahresdatensatz zusammen. Diesen nutzt das BAS zum einen selbst für die jährliche Fortschreibung des morbiditätsorientierten Risikostrukturausgleichs (Kap. 5.5.1). Zum anderen übermittelte das BAS diesen Gesamtdatensatz wieder an den Spitzenverband Bund der Krankenkassen (der diese Daten für eigene Planungs- und Entwicklungsaktivitäten nutzen darf) und an das Deutsche Institut für Medizinische Dokumentation und Information (DIMDI; inzwischen ein Teil des Bundesinstituts für Arzneimittel und Medizinprodukte – BfArM). Das DIMDI/BfArM soll als Datenstelle diese nationalen Versorgungsdatenbestände kontinuierlich fortschreiben, erweitern und für definierte Nutzungszwecke (§ 303 Abs. 2 SGB V) bereitstellen.

Zunächst wurden alle Datentreuhandelemente (Kap. 3.3.3) von zwei eigenständigen Organisationseinheiten des DIMDI realisiert. In der Datenannahmestelle wurde das versichertenbezogene Lieferpseudonym durch ein periodenübergreifendes Pseudonym ersetzt, das die jährliche Fortschreibung der einzelfallbezogenen Datensätze ermöglicht. Die Datenaufbereitungsstelle übernahm alle weiteren Aufgaben: Schrittweise wurden die Jahresdatensätze beginnend mit dem Jahr 2009 in die Data-Warehouse-Strukturen integriert und das »Informationssystem Versorgungsdaten« aufgebaut. Dazu wurden auch die jährlich fortgeschriebenen Diagnose-, Operationen-/Prozeduren- und Entgeltkataloge sowie relevante Angaben des Pharmazentralregisters in das Data Warehouse aufgenommen und mit den jeweiligen Versorgungsdaten verknüpft (Abb. 5.2). Zudem wurde eine Verknüpfung zum INKAR-Atlas des Bundesinstituts für Bau-, Stadt- und Raumforschung geschaffen. Dadurch können die Versorgungsdaten mit zahlreichen Indikatoren zu regionalen Lebensbedingungen in Deutschland auf Landkreisebene verknüpft werden, um z. B. den Einfluss sozioökonomischer Faktoren auf Gesundheit und Versorgung zu analysieren (DIMDI 2016, S. 7)

Abb. 5.2 Datenmodell: nationaler Versorgungsdatenbestand



DRG: Diagnosis Related Groups
EBM: Einheitlicher Bewertungsmaßstab
ICD: International Classification of Diseases
OPS: Operationen- und Prozedurenschlüssel
PZN: Pharmazentralnummer
PLZ: Postleitzahl

Quelle: § 303b SGB V; § 3 DaTraV

Die Entscheidung, die aufbereiteten Morbi-RSA-Daten für das Informationssystem Versorgungsdaten zu nutzen, führte aufgrund der langwierigen, mehrstufigen Aufbereitung und Weiterleitung in der Anfangsphase zu einem Zeitverzug von mehr als 4 Jahren bis zur Datenbereitstellung. Anfang 2014 startete der Pilotbetrieb der Datennutzung. Die Daten werden berechtigten Institutionen (ausschließlich Organe der Selbstverwaltung, Gesundheitsüberwachung und deren Aufsichtsgremien sowie öffentliche Forschungseinrichtungen) für definierte Zwecke (Planung, Weiterentwicklung der Versorgung, Gesundheitsberichterstattung sowie Forschung) auf Antrag und nach Prüfung in unterschiedlichen Anonymisierungsstufen zugänglich gemacht (§ 303e SGB V). Wird einem Untersuchungsantrag stattgegeben, stellt die Datenaufbereitungsstelle faktisch anonymisierte Datensätze als Scientific Use Files oder formal anonymisierte Datensätze für eine kontrollierte Datenfernverarbeitung bereit. Bei letzterer senden die jeweiligen Institutionen ihre Analyseskripte an die Mitarbeiter/innen der Datenaufbereitungsstelle, die diese Skripte ausführen, anschließend ermittelte Ergebnisse auf Anonymität prüfen, ggf. vergröbern und nur diese Ergebnisse an die jeweilige Institution übermitteln. Diesen Aufwand darf sich das DIMDI mittels Gebühren erstatten lassen, die das BMG per Verordnung festlegt (§ 303e SGB V).

Auch wenn die Datenbereitstellung grundsätzlich begrüßt wird, äußerten Fachkreise Kritik am etablierten Verfahren. Die erste interne Evaluation wurde für den Zeitraum Juli 2013 bis Februar 2016 vorgenommen (DIMDI 2016): In

diesen 32 Monaten wurden insgesamt lediglich 38 Bearbeitungsanträge gestellt – deutlich weniger als erwartet. Die Ursachen für die begrenzte Nachfrage seien strukturell bedingt. Zur Antragstellung berechtigte Institutionen und Nutzungszwecke lägen vorrangig im Bereich der gesundheitssystemischen Steuerung und Weiterentwicklung. Dafür sind die Daten jedoch meist nicht aktuell genug. Zudem könnten etliche Organe der Selbstverwaltung wie z. B. die Kassenärztliche Bundesvereinigung, das Deutsche Arzneiprüfungsinstitut oder große GKK Teildatenbestände zeitnäher analysieren und aktuellere Ergebnisse liefern. Der Nutzung im wissenschaftlichen Bereich sind sehr enge Grenzen gesetzt. Zum einen ist die Versorgungsforschung in unabhängigen wissenschaftlichen Einrichtungen nicht sehr weit verbreitet. Zum anderen dürfen die Daten laut Bundesbeauftragtem für den Datenschutz und Informationsfreiheit grundsätzlich nicht zu Ausarbeitungen für die Erlangung akademischer Grade (von Bachelor bis Habilitation) genutzt werden, da dies Privatangelegenheiten seien und die Verwendung für persönliche Zwecke ausgeschlossen ist. Im Beobachtungszeitraum betrug die mittlere Bearbeitungszeit pro Antrag 60 Stunden und somit zehnmal so lange, wie ursprünglich erwartet worden war. Die Ursachen dafür seien u. a., dass Fragestellungen und Methodik ohne detaillierte Kenntnisse der Daten erarbeitet wurden und die Datenanalyst/innen beim DIMDI sich in die inhaltliche Bearbeitung der Anträge aufwendig einarbeiten müssten, jedoch parallel noch andere Aufgaben hätten.

Im Rahmen der Evaluation wurde diverse Handlungsoptionen erarbeitet, um den Zeitverzug zu reduzieren, das Antragsverfahren schneller, transparenter und gerechter zu gestalten, die Gebühren aufwandsgerechter festzulegen und die Ressourcen für die inhaltliche Antragsbearbeitung zu erhöhen. Eine substanzielle Ausweitung der Nutzungsberechtigten und der Anwendungszwecke gehörte nicht dazu (es wurde lediglich empfohlen, die Daten zur Erlangung akademischer Grade nutzen zu dürfen).

Durch das 2019 verabschiedete Digitale-Versorgung-Gesetz (DVG) und die Neufassung der Datentransparenzverordnung in 2020 sollen einige der empfohlenen Veränderungen perspektivisch realisiert werden. Inhaltlich wurden die zu übermittelnden Datensätze erweitert (u. a. müssen zukünftig auch Angaben zum Vitalstatus und Sterbedatum übermittelt werden; Abb. 5.2). Um den Zeitraum bis zur Datenbereitstellung zu verkürzen, soll zukünftig bereits der Spitzenverband Bund der Krankenkassen als zentrale Datenannahmestelle die bei ihm zusammenlaufenden Daten selbst auf Vollständigkeit, Plausibilität und Konsistenz prüfen sowie Auffälligkeiten mit den jeweiligen GKK klären und dann ohne Umweg über das BAS den geprüften Jahresdatensatz direkt an die Vertrauensstelle übermitteln, die die finale periodenübergreifende Pseudonymisierung vornimmt. Diese Stelle wird organisatorisch und räumlich zum Robert Koch Institut verlegt. Die Datenaufbereitungsstelle wird zu einem Forschungs-

datenzentrum ausgebaut. Unterschiedliche Teams erhalten ein größeres Aufgabenspektrum, u. a. soll das Team, das Analyseanträge prüft, ein öffentliches Antragsregister mit Informationen zu den Antragstellenden, deren Vorhaben und den Ergebnissen aufbauen und pflegen. In den Nutzungslizenzen werden Reidentifizierungen explizit als Fehlverhalten benannt und mit einer Datenzugangssperre bis zu 2 Jahren belegt. Weitgehend unverändert blieb die ausschließliche Zugangsmöglichkeit für öffentliche Forschungsaktivitäten, was industrieseitig natürlich kritisiert wird (vfa 2020). Versichertenverbände kritisieren fehlende Einwilligungs- bzw. Widerspruchsmöglichkeiten für Betroffene sowie die gänzlich fehlende Beteiligung von Patienten-/Verbrauchervertretungen (z. B. BAG Selbsthilfe 2020).

Nach jetzigem Planungsstand soll dieses Forschungsdatenzentrum perspektivisch neben den nationalen Versorgungsdaten einen weiteren Datenbestand treuhänderisch verwalten: Die in der versichertengeführten elektronischen Patientenakte selbstverwalteten medizinischen Behandlungsdaten sowie individuell erhobene Gesundheitsdaten sollen Versicherte ab 2023 dem Forschungsdatenzentrum anonymisiert zur Weiterverwendung zu treuen Händen übergeben können (auch als Datenspende bezeichnet). Inwiefern Versicherte dies in relevantem Umfang tun werden, lässt sich derzeit nicht seriös abschätzen.

Einschätzung

Die großen Krankenkassen verfügen über einen erheblichen Bestand standardisierter Leistungsabrechnungs- bzw. Versorgungsdaten, die zeitnah zur Leistungsabrechnung über die kasseninternen Data Warehouses bereitgestellt werden können. Auch wenn diese Bestände lediglich Stichproben der nationalen Versorgungssituation darstellen, reichen diese oftmals aus, um Data-Mining-Projekte durchzuführen. Dafür benötigen die jeweiligen Krankenkassen zum einen spezielle wissenschaftliche Arbeitsgruppen, Abteilungen oder Institute und zum anderen die Zustimmung der jeweiligen Aufsichtsbehörde. Diese Datenanalyst/innen führen seit Jahren Data-Mining-Projekte durch. Teilweise werden die entwickelten Verfahren und Prozeduren verstetigt und liefern anhand aktualisierter Daten regelmäßig Informationen für unterschiedliche Entscheidungssituationen (z. B. zur Frage, welches Krankenhaus bei welchen Erkrankungen gute Behandlungsleistungen erzielt [Krankenhausnavigator]).

Zudem wird ein nationaler Versorgungsdatenbestand aufgebaut und fortgeschrieben – eine Totalerhebung, die im SGB V mit dem Begriff der Datentransparenz assoziiert wird. Aufgrund der geringen Datenaktualität und der eng begrenzten Nutzungsmöglichkeiten sind die bisherigen Analyseanfragen hinter den Erwartungen zurückgebliebenen. Auch mit der Neuordnung der diesbezüglichen Vorgehensweise und mit dem Ausbau des Betreibermodells zu einem Forschungsdatenzentrum werden Jahre vergehen, bis Versorgungsdaten über

das Informationssystem bereitgestellt werden. Zudem wurden die Nutzungsmöglichkeiten kaum erweitert. Zwei Forderungen der DSGVO in Bezug auf die privilegierte Datenweiterverwendung zu Forschungszwecken wurden auch mit der Neuordnung nicht umgesetzt, zum einen die bezüglich eines datenbezogenen Einwilligungsmanagements und zum andern die bezüglich einer weiten Auslegung des Forschungsbegriffs (Kap. 3.3.4). In gesundheitsbezogenen Datenstrategien und Innovationsinitiativen der Bundesregierung ist der Auf- und Ausbau dieses Forschungsdatenzentrums eine Schwerpunktaktivität im Bereich Gesundheit (Bundesregierung 2020a, S. 4, u. 2021a, S. 30). Die nächste Evaluation der Aktivitäten zur Datentransparenz ist für Ende 2023 vorgesehen. Die Entwicklung der Nachfrage sowie die Verwendung der aus komplexen Analysen resultierenden Ergebnisse wären dafür wichtige Punkte. Auch könnten sie mit dem Aufwand für die Datenbereitstellung abgewogen werden.

5.5.3 Sekundärnutzung von Leistungsdaten: Data-Mining-Beispiel »Pharmakovigilanz«

Jahre bevor der nationale Versorgungsdatenbestand aufgebaut wurde, begannen Überlegungen zwischen vier GKK und einer außeruniversitären Forschungseinrichtung, die bei den GKK zusammenlaufenden Daten aus der Leistungsabrechnung für die Sicherheitsüberwachung von Arzneimitteln unter Anwendungsbedingungen zu nutzen (Pharmakovigilanz). Rechtlich ist das im Rahmen begrenzter Forschungs- und Planungsvorhaben grundsätzlich möglich, wobei Aufsichtsbehörden u. a. prüfen, inwiefern das öffentliche Interesse an der Forschung das Geheimhaltungsinteresse betroffener Versicherter erheblich überwiegt und die schutzwürdigen Interessen Betroffener bei der Analyse angemessen gesichert werden (§ 287 SGB V; § 67c SGB X). Die Überwachung der Arzneimittelsicherheit ist zweifellos eine solche Aufgabe im öffentlichen Interesse. Durch das Arzneimittelgesetz ist sie bereits als kontinuierliche Aufgabe der jeweiligen Hersteller (zu Gewährleistung einer hohen Produktsicherheit), aber auch staatlicher Aufsichtsbehörden (im Rahmen der staatlichen Risikovorsorge) definiert. Die vier GKK und die Forschungseinrichtung haben die Hoffnung, dass diese etablierten Elemente zur Pharmakovigilanz durch Data-Mining mit Leistungsdaten ergänzt werden kann. Dazu soll nachfolgend die derzeitige Situation und der Data-Mining-Ansatz dargestellt und international verglichen und bewertet werden.

5.5.3.1 Pharmakovigilanz: die derzeitige Situation

Arzneimittel erhalten in allen Industrieländern erst dann eine Marktzulassung, wenn deren Sicherheit und Wirksamkeit durch klinische Studien nach international weitgehend einheitlichen Standards nachgewiesen sowie die Qualität der

Produktion belegt und geprüft wurden. Die Sicherheitsüberwachung nach der Marktzulassung baut wesentlich auf den Erkenntnissen aus den Sicherheitsprüfungen der Vorklinik und der klinischen Studien auf. In allen Studienphasen wird bei allen auftretenden gesundheitsrelevanten Problemen im Einzelfall medizinisch geprüft, ob die Einnahme des getesteten Arzneimittels eine Ursache dafür ist (Kausalitätsbeurteilung eines Verdachts auf unerwünschte Arzneimittelwirkung – UAW). Ergebnissen solcher Prüfungen im Rahmen klinischer Studien wird eine hohe Validität attestiert. Bestätigt sich dieser Verdacht, wird er in die UAW-Liste des Arzneimittels aufgenommen und geprüft, wie häufig und schwerwiegend diese unerwünschte Wirkung ist. Nur wenn der erwartete Nutzen des neuen Mittels das UAW-Risiko übertrifft, wird die Prüfung fortgesetzt und ggf. eine Zulassung erteilt.

Arzneimittelhersteller und studienrealisierende Ärzte/innen haften gemeinsam bei allen gesundheitlichen Schäden, die während der Studie bei Proband/innen entstehen. Nach der Zulassung haften sie bei auftretenden UAW nicht, wenn sie über alle ihnen bekannten UAW sachgerecht informieren (Packungsbeilage) und sorgfältig arbeiten (Kap. 4.2.3). Da Studien begrenzte Stichproben sind, können insbesondere UAW, die sehr selten, zeitlich verzögert oder nur bei einzelnen Risikogruppen auftreten (z. B. ältere Personen mit Komedikationen), nicht zuverlässig detektiert werden. Deshalb wird die Sicherheitsüberwachung auch nach der Zulassung fortgesetzt. Bisher stützt sich diese im Wesentlichen auf zwei Instrumente:

- Durch *Phase-IV-Studien* werden die (Neben-)Wirkungen von Arzneimitteln unter Anwendungsbedingungen weiter beobachtet. Diese Studien decken im Vergleich zu zulassungsrelevanten klinischen Studien meist längere Zeiträume ab, schließen meist größere Personengruppen ein und verfolgen oft mehrere Ziele gleichzeitig. Sie sollen sowohl die Sicherheit als auch den (Zusatz-)Nutzen eines Arzneimittels unter Anwendungsbedingungen belegen.
- In *Spontanmelderegistern* werden alle Meldungen zu UAW-Verdachtsfällen gesammelt. Solche Register werden einerseits bei Herstellern produktbezogen und andererseits produktübergreifend durch Monitoringstellen geführt. In Deutschland führt das Bundesinstitut für Arzneimittelsicherheit dieses nationale Register. Hersteller sind verpflichtet, alle ihnen angezeigten UAW-Verdachtsfälle zeitnah an das jeweilige nationale Register zu melden. Parallel gibt es über die jeweiligen Berufsordnungen der Ärzte/innen und Apotheker/innen eine Selbstverpflichtung, UAW-Verdachtsfälle an deren jeweilige Arzneimittelkommission zu melden, die diese Meldungen ebenfalls an nationale Register weiterleitet. Seit einigen Jahren können auch Einzelpersonen als Betroffene UAW-Verdachtsfälle dem BfArM direkt melden.

Mit diesen beiden Instrumenten können meist auch solche UAW aufgedeckt werden, die in klinischen Studien bis zur Zulassung unerkannt blieben. Jedoch haben beide Instrumente methodische Schwächen:

- In Phase-IV-Studien wird der Arzneimitteleinsatz lediglich beobachtet, ohne dass in die Behandlung eingegriffen wird. Infolgedessen fehlen oft spezifische Untersuchungen und Befunde, anhand derer UAW-Verdachtsfälle geprüft und sich Kausalitätsbeurteilungen fundieren lassen. Diese Beobachtungsstudien können wichtige Hinweise auf UAW liefern, jedoch wird je nach Sachlage im Einzelfall die Validität der Ergebnisse infrage gestellt, teilweise als nicht ausreichend bewertet und infolgedessen werden weitere konfirmatorische Studien gefordert, die zusätzliche externe Belege für die Richtigkeit der Ergebnisse liefern sollen (Kasten 5.1).
- Spontanmelderegister gelten keineswegs als umfassend. UAW, die erst mit erheblicher zeitlicher Verzögerung auftreten und/oder zu denen es noch nicht einmal Vermutungen gibt, werden durch Ärzte/innen, Apotheker/innen und/oder Patient/innen kaum erkannt. Die Medienaufmerksamkeit beeinflusst das Meldeverhalten erheblich: Wird über UAW-Vermutungen berichtet, steigt die Zahl der Meldungen deutlich. Bisher werden UAW-Verdachtsfälle überwiegend von Pharmafirmen gemeldet (ca. 85 % laut BfArM 2016). Von medizinischem Personal oder Betroffenen gibt es weit weniger Meldungen. Durch die unterschiedlichen Meldekanäle sind Mehrfachnennungen nicht unwahrscheinlich. Oftmals gibt es nur lückenhafte Angaben zum Gesundheitszustand der/des Betroffenen, sodass eine Kausalitätsbeurteilung der UAW-Verdachtsfälle nur eingeschränkt möglich ist. Mögliche UAW-Risiken lassen sich aus Spontanmelderegistern nicht quantifizieren.

Mit diesen Instrumenten dauerte es laut Ware (2005) im Mittel 5 Jahre, bis ausreichend Belege für die Revision einer Nutzen-Risiko-Abwägung zu einem Arzneimittel vorlagen. Methodische Schwächen gibt es vor allem bei der Erkennung bislang völlig unbekannter UAW-Signale und bei der Abschätzung der gesundheitlichen Relevanz auf der Grundlage relativer Häufigkeiten (Risikoquantifizierung z. B. selten oder sehr seltenes Auftreten).

Gesundheitsexperten betonen seit Jahren, dass systematischere Analysen von Daten aus der Behandlung (Real-Life-Daten) die derzeitigen Pharmakovigilanzinstrumente ergänzen und verbessern können. Die US-amerikanische Food and Drug Administration (FDA) hat beim Einsatz und der Überwachung von Vioxx© bereits vor mehr als 10 Jahren die Tür geöffnet für Analysen, die auf umfassenderen Datenbeständen aus der medizinischen Versorgung aufbauen (Kasten 5.1). Auch in Deutschland wird seit einigen Jahren getestet, inwiefern sich durch die Analyse von Daten aus der medizinischen Versorgung zusätzliche Informationen zu möglichen UAW generieren lassen.

Kasten 5.1 Data-Mining im Rahmen der Pharmakovigilanz (Beispiele)

Wirkstoff: Rofecoxib (Handelsname: Vioxx©)

Der Wirkstoff wurde entwickelt, um Schmerzen selektiv zu bekämpfen, ohne die Magenschleimhäute anzugreifen (Zusatznutzen gegenüber verfügbaren Arzneimittel, die ein erhöhtes Risiko für gastrointestinale Blutungen hatten). Vioxx© wurde 1999 in den USA und in Europa zur Behandlung rheumatischer Erkrankungen und Schmerzen zugelassen. Innerhalb von 5 Jahren wurden ca. 80 Mio. Menschen damit behandelt. Die Jahresumsätze überstiegen 2 Mrd. US-Dollar – ein Blockbuster für den Hersteller.

Mit der Zulassung startete der Hersteller u. a. eine Phase-IV-Studie, in der das Wirkungsspektrum von Vioxx© mit einem anderen Schmerzmittel verglichen wurde. Anhand der Studienergebnisse wurde einerseits der Zusatznutzen gegenüber anderen Schmerzmitteln bestätigt, andererseits wurden bereits Ende 2000 erste Sicherheitsbedenken geäußert: Bei Patient/innen mit Herz-Kreislauf-Problemen würde durch Vioxx© das Herzinfarktrisiko steigen. Der Hersteller vertrat die These, dass nicht Vioxx© häufiger solche Probleme verursacht, sondern vielmehr das Vergleichspräparat tendenziell herzschonender sei und deshalb dort unterdurchschnittlich wenige Herz-Kreislauf-Probleme auftreten würden. Es folgten weitere Studien, die je nach Studiendesign die Hypothese erhöhter kardiovaskulärer Risiken belegten oder nicht belegten. Aufgrund dieser Studienergebnisse forderte die FDA 2002 den Hersteller lediglich auf, in Fach- und Patienteninformationen stärker auf Herz-Kreislauf-Erkrankungen als mögliche UAW hinzuweisen. Um die Beweislage eindeutiger zu klären, beschritt die FDA 2003/2004 einen neuen Weg: Sie finanzierte ein Data-Mining-Projekt, das eine große US-amerikanische Health Maintenance Organization (HMO) realisierte. HMOs sind Krankenversicherungen mit eigenen Kliniken und Praxen, die medizinischen Behandlungsdaten für alle bei ihnen versicherten Personen zentral verwalten. Diese medizinischen Dokumentationen der Versichertenpopulation ist weit umfassender als die einer Phase-IV-Studienpopulation. Die Analyse ergab ein 1,5-mal größeres Herzinfarkt-/Herztodrisiko bei niedriger Vioxx©-Dosierung im Vergleich zu Behandlungen mit anderen Schmerzmitteln, bei höherer Dosierung verdreifachte sich das Risiko. Die FDA rechnete die Zahlen hoch und schätzte, dass von 1999 bis 2003 88.000 bis 134.000 Herz-Kreislauf-Erkrankungen auf die Einnahme von Vioxx® zurückzuführen seien, von denen 30 bis 40 % tödlich waren (Graham et al. 2005) – ein Beleg für die bereits im Jahr 2000 geäußerten Risikobedenken, für die zeitgleich auch biologische Erklärungen geliefert wurden. Der Hersteller rief sämtliche Tabletten zurück und nahm die Substanz umgehend vom Markt. Das Medienecho war riesig (ausführlich z. B. in Ware 2005). Jahrelange juristische

Auseinandersetzungen zu Haftung und Schadensersatz folgten. Der US-amerikanische Kongress befasste sich mit der Thematik und forderte die FDA auf, das bestehende Pharmakovigilanzsystem zu verbessern. Dies war der Auslöser der »Sentinel Initiative«.

Wirkstoff: Rosigliatzon (Handelsname: Avandia®)

Der Wirkstoff zielt darauf ab, die körpereigene Insulinsensitivität zu verbessern, um der nachlassenden Wirksamkeit körpereigenen Insulins entgegenzuwirken. Er wurde 1999 in den USA und Europa zur Behandlung von Typ-2-Diabetes zugelassen. Eine herstellerfinanzierte Phase-IV-Studie startete in mehreren europäischen Ländern, um Langzeitnebenwirkungen, darunter Herz-Kreislauf-Probleme, zu überwachen. Aufgrund der großen Verbreitung von Typ-2-Diabetes und der erforderlichen Dauerbehandlung wurde die patentgeschützte Substanz ebenfalls ein Blockbuster (Jahresumsatz 2006: 2,5 bis 3 Mrd. US-Dollar).

2007 führte ein US-amerikanischer Kardiologe mit Kollegen eine Metaanalyse von 42 publizierten Studien zu den Folgen der Substanzverwendung durch. Mit ihrem Data-Mining-Ansatz ermittelten sie ein um über 40 % erhöhtes Herzinfarktrisiko bei Langzeitnutzenden (Nissen/Wolski 2007). Der Hersteller verwies auf die von ihm finanzierte Phase-IV-Studie, die keinen Beleg für ein erhöhtes Herzinfarktrisiko lieferte (Home et al. 2009). Nach kontroverser Methodendiskussion empfahl die EMA 2010 die Marktzulassung auszusetzen (GSK 2010). Die FDA schränkte die Verwendung der Substanz ein (Zurückstufung als Reservemedikament) und ordnete eine strenge Risikoüberwachung an und beauftragte das Duke Clinical Research Institute mit der Überprüfung der Studienergebnisse. Auch dieser Fall verursachte ein großes Medienecho. Die Umsätze brachen sofort weltweit ein. Als 2012 das Patent auslief, wurden nur noch wenige Tausend Patient/innen in den USA damit behandelt, der Jahresumsatz lag im einstelligen Millionenbereich. 2013 legte das Duke Clinical Research Institute seine Ergebnisse vor. Die Revision der Studiendaten und nochmalige Analysen lieferten keine Belege für ein erhöhtes Herzinfarktrisiko. Zwar lockerte die FDA daraufhin die Risikoüberwachung, dennoch rechnen Marktbeobachter nicht damit, dass diese Substanz als Antidiabetikum in den Industrieländern wieder relevant wird.

5.5.3.2 Die pharmakoepidemiologische Forschungsdatenbank

Im Jahr 2004, nach dem Vioxx-Skandal und der FDA-HMO-Kooperation (Kasten 5.1) sowie im Zuge der 2003 eingeleiteten GKV-Modernisierung, die GKK mit umfangreicheren Leistungsdatensätzen ausstattete, beantragten vier GKK und das Bremer Leibniz-Institut für Präventionsforschung und Epidemiologie

(BIPS) in Kooperation mit der Universität Bremen den Aufbau der pharmakoepidemiologischen Forschungsdatenbank (»German Pharmacoepidemiological Research Database« – GePaRD) (Schepers et al. 2015, S. 82 ff.). In Anlehnung an das FDA-HMO-Studiendesign wurde das GePaRD-Datenmodell auf der Grundlage der Leistungsabrechnungsdaten und entsprechend der nationalen Datenschutzvorgaben das Treuhandverfahren entwickelt und mit den zuständigen Versicherungsaufsichts- und Datenschutzgremien abgestimmt. Die vier beteiligten GKK übermitteln von ihren ca. 20 Mio. bundesweit Versicherten neben ausgewählten pseudonymisierten Versichertenstammdaten, die Diagnose- und Behandlungsdaten mit deren Zeitangaben aus der ambulanten und stationären Leistungsabrechnung sowie die arzneimittelbezogenen Daten mit Zeitangaben von den Rezeptabrechnungen, jedoch keinerlei Kostenpositionen. Zudem werden die Kataloge der Diagnose- und Leistungscodierung und die Daten des nationalen Pharmazentralregisters (mit Angaben zu Packungsgrößen, Tagesdosierung, Wirkstoffbestandteilen etc.; Kap. 5.4.1) integriert, nicht aber andere Heil- und Hilfsmittelleistungen oder ergänzende sozio-ökonomische Faktoren. Statt bei jedem Versicherten die Einwilligung zur spezifischen sekundären Nutzung zu Forschungszwecken einzuholen – wofür es bisher ohnehin keine standardisierten Verfahren gibt – wurde folgendes Datentreuhandverfahren vereinbart (Schepers et al. 2015, S. 88):

1. Die GKK stellen definierte Leistungsdatensätze jahrgangsweise zusammen, pseudonymisieren die KV-Nr. der Versicherten anhand eines kassenintern en Schlüssels und übermitteln die Datensätze an eine Vertrauensstelle an der Universität Bremen.
2. Die Vertrauensstelle prüft die Datensätze mit Routinealgorithmen, klärt Unplausibilitäten mit der jeweiligen GKK und leitet aus dem kassenintern en Schlüssel ein periodenübergreifendes Pseudonym ab (Zweifachpseudonymisierung). Von den Krankenkassen gelieferte Daten, der interne Schlüssel und die aufbereiteten Jahresdaten werden getrennt gespeichert. Nur die periodenübergreifend pseudonymisierten Jahresdatensätze werden an die GePaRD-Registerstelle weitergegeben, die diese über das Pseudonym an den existieren Datenbestand anfügt und diesen fortschreibt. Die Registerstelle integriert auch die jährlich aktualisierten Codierungskataloge und das Pharmazentralregister in die pharmakoepidemiologische Forschungsdatenbank.
3. Nur BIPS-Mitarbeitende dürfen eine konkrete Datennutzung zum einen bei den beteiligten GKK und zum anderen bei den zuständigen Aufsichtsbehörden beantragen. Sie definieren die Untersuchungsfrage und die für die Analyse erforderliche Datenmenge. Die GKK und Aufsichtsbehörden prüfen separat und genehmigen gegebenenfalls. Externe Wissenschaftler/innen können sich über Kooperationen beteiligen.

4. Entsprechend der jeweiligen Genehmigungen extrahiert die Registerstelle aus dem GePaRD-Datenbestand den definierten Analysedatensatz und stellt diesen dem/der verantwortlichen BIPS-Wissenschaftler/in bereit.
5. Verantwortliche Wissenschaftler/innen führen die jeweiligen Analysen eigenverantwortlich mit ihrem Team durch, prüfen die Ergebnisse und veröffentlichen sie gegebenenfalls.

Das GePaRD-Verfahren führt zu einer faktischen Anonymisierung der Leistungsdaten. Dadurch können das GKK-Verbot zur sektorübergreifenden Datenzusammenführung und die gesetzlich definierten Löschfristen überwunden werden (für Forschungszwecke). Datenprüfung, Aufbereitung und Bereitstellung führen zu einer zeitlichen Verzögerung von ca. 2 Jahren. Der GePaRD-Datenbestand ermöglicht vielfältige Analysen zum Einsatz von Arzneimitteln und dessen Folgen. Er hat inhaltlich erhebliche Schnittmengen zum nationalen Versorgungsdatenbestand (Kap. 4.5.2), ist jedoch nicht so umfangreich (Stichprobe statt Vollerhebung; kein Heil- und Hilfsmitteleinsatz), dafür aktueller und deckt zudem einen längeren Zeitraum ab. Die GePaRD-Forschungsdatenbank wurde 2020 vom Rat für Sozial- und Wirtschaftsdaten vorläufig akkreditiert und wird inzwischen als ein Zentrum der Forschungsdateninfrastruktur (FDI) geführt (Kap. 3.3.3). Über die FDI sollen verfügbare Datenbestände sichtbarer gemacht und die Nutzung entsprechend der jeweiligen Rechtsstruktur verbessert werden. Damit könnten über die FDI Kooperationsanfragen an das BIPS gestellt werden. Die praktische Relevanz dieser neuen Option lässt sich bisher nicht abschätzen.

5.5.3.3 Der Data-Mining-Prozess

Datenbasis zur Suche nach unerwünschten Arzneimittelwirkungen

Für die datenbasierte Suche nach möglichen unerwünschten Arzneimittelwirkungen (UAW) werden einerseits Daten zur Arzneimitteleinnahme und andererseits Daten zu gesundheitlichen Beeinträchtigungen benötigt. Eine Ursache-Wirkungs-Beziehung kann hypothetisch nur bei Beeinträchtigungen unterstellt werden, die zeitlich nach einer Arzneimitteleinnahme aufgetreten sind. Deshalb sind die Zeitangaben von besonderer Relevanz.

- *Arzneimitteleinnahme*: Es gibt in Deutschland gegenwärtig keine systematische Erfassung der personenbezogenen Arzneimitteleinnahme. Im stationären Bereich sollte diese dokumentiert werden, jedoch unterliegen die arztgeführten Primärakten der Schweigepflicht. Digitale, zu Forschungszwecken nutzbare Medikationspläne gibt es bisher nicht. Eine Näherung ist über die versichertenbezogenen Rezeptdaten der Apotheken möglich, auch wenn diese Daten nur die ambulante Abgabe rezeptpflichtiger Arzneimittel enthalten, deren Kosten die jeweilige GKK teilweise trägt. Für die GePaRD-

Datenbank werden die Rezeptangaben mithilfe des nationalen Pharmazentralregisters (Kap. 5.4.1) teilweise umcodiert und erweitert, sodass für derzeit ca. 6.500 Wirkstoffe bzw. Wirkstoffkombinationen der jeweilige Einnahmezeitraum und die Dosierung näherungsweise geschätzt werden können.

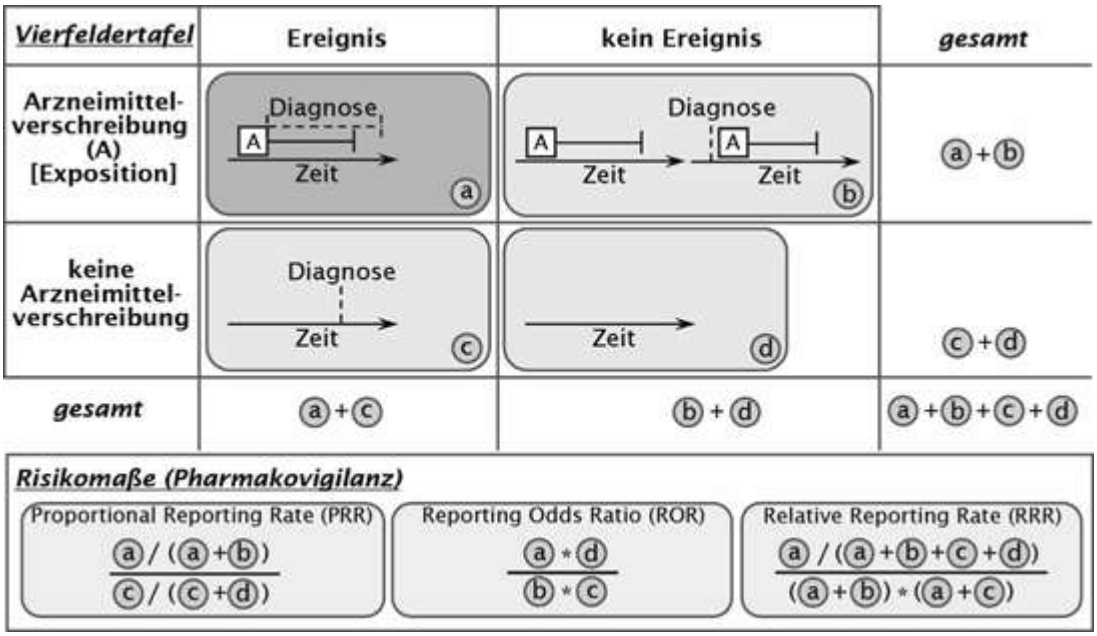
- *Gesundheitliche Beeinträchtigungen*: GKK erhalten seit 2004 mit den Leistungsabrechnungsdaten u. a. codierte Haupt- und Nebendiagnosen von allen Kliniken und Praxen. Die ICD-Klassifikation, mit der die Diagnosen für die Leistungsabrechnung codiert werden, wird jährlich überarbeitet und an medizinische Entwicklungen angepasst. Insbesondere für Untersuchungen über längere Zeiträume (Längsschnittanalysen) sind diese Überarbeitungen z. T. problematisch, da sie zu Brüchen in den Datenbeständen im Zeitverlauf führen können. Zwar sind die medizinischen Einrichtungen zur korrekten Abrechnung verpflichtet und die Richtigkeit und Validität der Angaben wird automatisiert geprüft, dennoch lässt sich nur bedingt abschätzen, inwiefern die zur Leistungsabrechnung übermittelten Haupt- und Nebendiagnosen gesundheitliche Probleme von Versicherten vollumfänglich und korrekt abbilden (medizinische Einrichtungen übermitteln i. d. R. nur abrechnungsrelevante Diagnosen; Kap. 4.5.1). Für die retrospektive datenbasierte UAW-Suche werden nur Diagnosen berücksichtigt, die während oder nach einer Arzneimittelabgabe auftraten. Deshalb sind die Zeitpunkte der Arzneimittelabgabe und der Diagnosen nötig.

Methodisches Vorgehen

Das Data-Mining-Verfahren basiert auf Disproportionalitätsanalysen von Vierfeldertafeln Abb. 5.3). Für jede in Betracht zu ziehende Arzneimittel-Diagnose-Kombination wird eine solche Tafel erstellt. Je umfangreicher der Analysedatensatz bzw. je größer die Stichprobe, desto zuverlässiger können die Werte aller vier Felder bestimmt und unterschiedliche klassische Risikomaße für jede Arzneimittel-Diagnose-Kombination berechnet werden (Abb. 5.3 unten). Diese Risikomaße können als Indikator für die Stärke eines UAW-Signals interpretiert werden.

Methodisch kommen klassische Risikomaße an ihre Grenzen, wenn unerwünschte Arzneimittelwirkungen nur sehr selten auftreten (dann sind hohe Werte kein zuverlässiger Indikator für ein UAW-Signal mehr). Soll geprüft werden, ob ein Arzneimittel möglicherweise in sehr seltenen Fällen unerwünschte Wirkungen hat, können inzwischen auch Bayes'sche Netze trainiert werden. Sie basieren auf bedingten Wahrscheinlichkeitsverteilungen und sind vergleichsweise rechenintensiv, können aber bei sehr selten auftretenden unerwünschten Arzneimittelwirkungen zuverlässige Signale generieren.

Abb. 5.3 Vierfeldertafel zum Finden unerwünschter Arzneimittelwirkungen



PRR: proportionale Meldequote; ROR: Meldung der Quotenverhältnisse; RRR: relative Meldefrequenz

Quelle: nach Schepers et al. 2015, S. 89

Mit den inzwischen verfügbaren Algorithmen könnten in umfangreichen Datenbeständen UAW-Signale auch sehr breit gesucht werden, d. h., es müssen wenig bis keine Restriktionen vorgegeben werden, auf welche Arzneimittel und auf welche Diagnosen sich die Suche beschränken soll. Die Anzahl dieser Tafeln wird dann jedoch schnell sehr groß. Mit den GePaRD-Datenbeständen könnten theoretisch bis zu 6.500 ATC-codierte pharmakologische Wirkstoffe/Wirkstoffkombinationen und ca. 12.000 ICD-codierte Diagnosen zu knapp 80 Mio. Vierfeldertafeln verknüpft werden (werden Kombinationen aus Haupt- und Nebendiagnosen berücksichtigt, kann sich die Zahl der Tafeln weiter erhöhen). Würden noch dazu sehr lange Zeiträume zwischen Arzneimittelverschreibung und möglicher unerwünschter Wirkung betrachtet und unbeschränkt in großen Datenbeständen gesucht, müsste damit gerechnet werden, dass sehr viele UAW-Signale generiert würden.

Ergebnisbewertung

Alle mit Disproportionalitätsanalysen ermittelten Ergebnisse werden bewertet und klassifiziert in

- bereits bekannte UAW-Signale (in nationalen und/oder internationalen Melderegistern enthalten und in den Bulletins zu Arzneimittelsicherheit bereits aufgeführt),
- aus medizinisch-fachlicher Sicht sehr unplausible Zusammenhänge sowie
- potenziell relevante UAW-Signale (die weiter beobachtet oder vertieft geprüft werden sollten).

Eine blinde oder anlasslose Suche ist einerseits methodisch wenig zielführend, da zu viele UAW-Signale generiert werden würden, auch längst bekannte sowie medizinisch-fachlich unplausible, die im Anschluss einzeln extern auf der Grundlage medizinischen Wissens bewertet werden müssen (Antes/Bertram 2019). Andererseits werden anlasslose Analysen normativ begrenzt, denn Analysen müssen einzeln bei den GKK und den Aufsichtsbehörden beantragt werden. Laut Schepers et al. (2015, S. 91) müssten sie dem Prinzip der Zweckbindung gemäß § 75 SGB X entsprechen und folglich müsse vorab spezifiziert werden, bei welchen Wirkstoffen/Wirkstoffkombinationen nach welchen unerwünschten Wirkungen gesucht werden soll. Man müsse zumindest UAW-Vermutungen haben und die Suche nach Auffälligkeiten in den Analysedaten eingrenzen können.

Laut Schepers et al. (2015, S. 92 f.) zielen die Data-Mining-Verfahren mittels GePaRD-Daten nicht darauf ab, UAW-Risiken eines Arzneimittels sicher festzustellen (das dafür nötige Evidenzniveau könne mit derartigen Sekundäranalysen von Leistungsdaten nicht erreicht werden). Sie erlauben lediglich, potenziell relevante UAW-Signale zu ermitteln, bei denen dann diskutiert werden müsse, welche weiteren Maßnahmen der Risikoüberwachung folgen sollen. Wenn sich ein potenzielles Sicherheitsrisiko andeutet, das UAW-Signal aber nicht sehr stark bzw. statistisch stabil ist, könne die Analyse mit anderen Datensätzen wiederholt werden (sofern nur ein Teil der GePaRD-Daten analysiert wurde, könnte der andere Teil zur Signalprüfung genutzt werden). Schwache Signale sollten weiter beobachtet werden. Bei deutlichen UAW-Signalen, sollte eine spezifische konfirmatorische Studie (Phase-IV-Studie) durchgeführt werden, um das Data-Mining-Ergebnis extern zu prüfen. Einen allgemeinen methodisch-fachlichen Konsens zur Bewertung der Ergebnisgüte von UAW-Signalen und zur Ableitung von Folgemaßnahmen gibt es nicht. Vielmehr wird im Einzelfall unter Beachtung des jeweiligen Kontextes entschieden (wie gravierend sind die Wirkungen im Vergleich zum Substanznutzen, gibt es Behandlungsalternativen).

5.5.3.4 Vorgehen im internationalen Vergleich

Mit dem Aufbau der GePaRD-Forschungsdatenbank und dem vorgestellten Verfahren, Leistungsabrechnungsdaten für Pharmakovigilanzanalysen zu nutzen, wurde 2004 in Deutschland Neuland betreten. Da aufgrund der schwachen

Anonymisierung der Daten (Kap. 3.3.3) ausschließlich BIPS-Mitarbeitende Analysevorhaben beantragen und federführend durchführen können, ist die Intensität der Datennutzung an die personellen Ressourcen des BIPS gebunden. Wie häufig auf die Daten zugegriffen wird, wie viele UAW-Signale bereits gefunden oder erhärtet und welche zusätzlichen Informationen generiert wurden, lässt sich von außen bisher nicht beurteilen. Systematische Dokumentationen durchgeführter Analysen und deren Ergebnisse, Verzeichnisse daraus hervorgegangener Publikationen oder Evaluationen der Datennutzung werden über die Internetseiten der Betreiberinstitution nicht bereitgestellt. In der Datenbankbeschreibung werden für die letzten Jahre 7 bis 12 wissenschaftliche Publikationen pro Jahr gelistet (BIPS 2017). Schepers et al. (2015, S. 95) verweisen beispielhaft auf eine Datenanalyse mit der ein Verdacht zur Verdopplung des Fieberkrampfrisikos bei einer Vierfachimpfkombination (Mumps, Masern, Röteln und Windpocken) im Vergleich zur vorher üblichen Dreifachkombination und separater Windpockenimpfung 2014 erhärtet wurde (Schink et al. 2014). Formuliert wurde dieser Verdacht in den USA bereits 2009 – 5 Jahre früher. Die Diskussion um den Zusatznutzen der Vierfachimpfung hält an. Marktverfügbar sind beide Impfstoffvarianten. Die ständige Impfkommission empfiehlt eine Impfung, äußert sich aber nicht zur Kombinationsform.

Im internationalen Vergleich ist der Ansatz, Daten, die im Rahmen der medizinischen Versorgung entstehen, kontinuierlich zusammenzuführen und zur Beantwortung unterschiedlicher Untersuchungsfragen sekundär zu nutzen, keinesfalls einzigartig. Auch in etlichen anderen Industrieländern werden seit Jahren Real-World Data für Forschungszwecke im Allgemeinen und für Pharmakovigilanzuntersuchungen im Besonderen zugänglich gemacht (Schepers et al. 2015, S. 94). Dabei sollten die Besonderheiten von medizinischen Behandlungsdaten und administrativen, vergütungsrelevanten Daten ggf. berücksichtigt werden.

Die US-amerikanische »Sentinel Initiative«

In den USA stand 2004 nach dem Vioxx®-Skandal nicht nur der Arzneimittelhersteller, sondern auch die Arzneimittelaufsichtsbehörde FDA und deren Risikoüberwachung in der Kritik (Kaplan 2017). Die FDA würde über keine ausreichenden herstellerunabhängigen Pharmakovigilanzverfahren verfügen und dadurch zu zögerlich reagieren, vor allem aber nicht proaktiv agieren können (Kasten 5.1). Die Situation wurde auf höchster Regierungsebene diskutiert. 2007 verabschiedete der Kongress den Food and Drug Administration Amendments Act (FDAAA), der die FDA u. a. aufforderte, ein besseres UAW-Frühwarnsystem einzurichten. Die »Sentinel Initiative« wurde als zusätzliches Element der Risikoüberwachung gesetzlich verankert. Unterschiedliche mögliche

Datenmodelle wurden diskutiert und gegeneinander abgewogen. Sie mündeten in vier Empfehlungen (Brown et al. 2009):

- *Datenbasis*: Zusätzlich zu den existierenden Monitoringelementen sollten Daten von einer erheblich größeren Population die Basis eines Frühwarnsystems bilden (100 Mio. Patient/innen wurden anvisiert, knapp ein Drittel der US-amerikanischen Bevölkerung). Unterschiedliche Kooperationen mit medizinischen Einrichtungen, um auf deren uneinheitlich codierte Behandlungsdaten zugreifen zu können, schienen nicht praktikabel. Es wurde empfohlen, administrative Daten, die ohnehin schon bei diversen Versicherungen zusammenliefen, als Datenbasis zu nehmen.
- *Datenverantwortung*: Die kooperierenden datenhaltenden Organisationen sollten als Data Owner die Kontrolle über die Datenhaltung und -verwendung behalten und für die Einhaltung von Datenschutz- und -sicherheit verantwortlich bleiben (an der »Sentinel Initiative« beteiligen sich inzwischen 19 Data Owner, vor allem Krankenversicherungen oder HMO).
- *Datenstruktur*: Die datenhaltenden Einrichtungen sollten gleiche Datenbankmodelle nutzen und einheitliche Bezeichnungen und Klassifikationen verwenden.
- *Datenanalysen*: Für einzelne Untersuchungsfragen sollten einheitliche Programmcodes zentral entwickelt und dezentral ausgeführt werden. Die datenhaltenden Institutionen melden lediglich Ergebnisse zurück (als Teil eines Privacy-Preserving-Konzepts; Kap. 3.3.3).

Das von Protagonisten als Quantensprung für die Pharmakovigilanz angekündigte Pilotprogramm begann 2009 in Kooperation mit dem Nonprofit-Healthservice-Unternehmen Harvard Pilgrim (Kaplan 2017). Der Aufbau der Grundstrukturen kostete in den ersten 7 Jahren mehr als 200 Mio. US-Dollar. Seit 2016 läuft der Regelbetrieb der »Sentinel Initiative«, durch die (theoretisch) Daten von insgesamt 193 Mio. Krankenversicherten genutzt werden könnten (60 % der Einwohner/innen). Um einen UAW-Verdacht zu erkennen, nutzt die FDA nach wie vor zunächst die klassischen Pharmakovigilanzinstrumente (Spontanmelderegister, Studienergebnisse). Wenn diese Hinweise zu möglichen UAWs liefern, beginnt nicht mehr die Abwägung, ob eine aufwendige und langwierige Phase-IV-Studie nötig ist, vielmehr beauftragt die FDA Harvard Pilgrim sofort mit einer Prüfung. Das Unternehmen beauftragt einen oder mehrere Data Owner, ihre Datensätze zu prüfen. Seit dem Start der »Sentinel-Initiative« 2009 wurden Hunderte Abfragen durchgeführt. Die Ergebnisse wurden oft innerhalb von Wochen vorgelegt. Wie oft danach Phase-IV-Studien angeordnet wurden oder inwiefern Hinweise nicht erhärtet wurden, ist dem TAB nicht bekannt. Die FDA verweist darauf, dass die Hinweise aus den Data-Mining-Verfahren in zwei Fällen dazu führten, die Informationen der Packungsbeilage zu ändern (Kaplan 2017).

Spätestens seit der Zwischenevaluation 2015 sind die euphorischen Ankündigungen einer kritischeren Auseinandersetzung mit den Grenzen des Verfahrens gewichen (FDA 2017). Im Zentrum der Diskussion steht die Datenbasis, also die administrativen Daten der Leistungsabrechnung bei Versicherungsgesellschaften. Darin wären lediglich Arzneimittelverschreibungen, aber keine Angaben enthalten, inwiefern die Mittel tatsächlich genommen wurden. Mögliche unerwünschte Wirkungen können nur über die Diagnosen erfasst, nicht aber durch Laborwerte oder andere medizinische Befunde fundiert werden. Bezüglich der Diagnoseerfassung gebe es unterschiedliche Gewohnheiten, oftmals werden nur abrechnungsrelevante Diagnosen, nicht aber klassische Nebenwirkungen wie Verdauungsprobleme, Konzentrationsmangel usw. erfasst. Kritiker vermuten, dass viele Menschen, die an den Folgen von Medikamenteneinnahmen starben, durch diese Datenbasis nicht erkannt werden könnten, denn der Tod und Todesursachen sind in den administrativen Daten zur Leistungsabrechnung nicht enthalten. Dadurch sei man sich unsicher, welche Fragen mit der »Sentinel-Initiative« zuverlässig beantwortet werden können. Man könne nicht zuverlässig abschätzen, ob mit dem Ansatz tatsächlich die Nebenwirkungen erkannt werden, nach denen man sucht und die man mit den gängigen Verfahren nur schlecht fand (Kaplan 2017). Für die strategische Weiterentwicklung bis 2023 soll neben daten- und informationstechnischen Aktivitäten auch die methodisch-analytische Auseinandersetzung fortgesetzt werden. Das »Sentinel-Ökosystem« als zusätzliche nationale Ressource gilt nach vor als Vision.[163]

Die britische Forschungsdatenbank »Clinical Practice Research Datalink«

Großbritannien gehört zu den Vorreitern der Zusammenführung von medizinischen Behandlungsdaten auf Einzelfallebene und deren Bereitstellung für wissenschaftliche Zwecke. Der heute verfügbare »Clinical Practice Research Datalink« (CPRD)[164] geht auf das Engagement eines niedergelassenen englischen Hausarztes und eines IT-Spezialisten zurück, die bereits in den 1980er Jahren aus longitudinalen Behandlungsdaten Informationen zum Wirkungsspektrum medizinischer Produkte unter Anwendungsbedingungen generieren wollten. Nahezu zeitgleich begannen nationale Aktivitäten, eine allgemeine Systematik medizinischer Begriffe auf- und auszubauen (READ-Code). Die Initiatoren entwickelten ein Modul für ein IT-Praxissystem, mit dem aus den bei der Behandlung kontinuierlich entstehenden Patientendaten definierte, codierte Merkmale automatisiert in eine eigenständige Forschungsdatenbank übertragen werden konnten.

163 www.fda.gov/safety/fdas-sentinel-initiative/fdas-sentinel-initiative-background (13.121.2021)
164 www.cprd.com (13.12.2021)

Diese individuell gestarteten Datensammel- und Analyseaktivitäten wurden in den 1990er Jahren in die Verantwortung des Gesundheitsministeriums gegeben, das den nationalen Gesundheitsdienst (National Health Service – NHS) und die Arzneimittelaufsicht mit der Fortentwicklung der Datenbank beauftragte. Seit 2007 gibt es eine strategische Allianz mit einem Marktforschungs- und Beratungsunternehmen, das seine Expertise bezüglich Gesundheitsmarktanalysen einbringt. 2011 startete eine Regierungsinitiative,[165] die durch die Intensivierung der forschungsseitigen Nutzung von »electronic health records« – standardisierte Teile medizinischer Akten, keine administrativen Daten zur Leistungsabrechnung – eine international herausgehobene Stellung im Gesundheitsforschungsbereich anstrebt. Inzwischen übermitteln mehr als 2.000 primärversorgende Einrichtungen regelmäßig definierte Behandlungsdaten (Diagnosen einschließlich Labordaten, medizinische Behandlungen einschließlich Arzneimittelverschreibungen) von ihren Patient/innen an die CPRD-Forschungsdatenbank, sofern Patient/innen dies nicht verbieten (Opt-out-Modell). Die Daten werden periodenübergreifend pseudonymisiert. Dadurch kann der CPRD-Bestand regelmäßig fortgeschrieben werden. Inzwischen enthält diese Datenbank Behandlungsdaten von insgesamt 60 Mio. Patient/innen. 16 Mio. Fälle gelten als aktiv, d. h., sie werden bis heute kontinuierlich fortgeschrieben und bilden teilweise eine mehr als 20-jährige Krankheitsgeschichte ab. Die Daten werden über einen Nonprofit-Datentreuhänder Wissenschaftler/innen mittels nichtübertragbarer Nutzungslizenzen für Forschungsaktivitäten mit einem potenziellen gesellschaftlichen Nutzen bereitgestellt. Die Nutzungsgebühren liegen zwischen 75.000 und 330.000 Britische Pfund pro Jahr je nach Nutzungsform.[166] Jegliche Reidentifizierung dieser lediglich formal anonymisierten Mikrodaten ist explizit untersagt. Der Treuhänder behält sich ein Kontrollrecht vor. Ein CPRD-Analyseteam unterstützt Wissenschaftler/innen bei ihren Aktivitäten. Die Datenbank wird kontinuierlich zu einem wissenschaftlichen Analyse- und Informationssystem ausgebaut mit vielfältigen vorinstallierten visuellen und analytischen Data-Mining-Funktionen. CPRD-Daten flossen bereits in knapp 3.000 begutachtete wissenschaftliche Publikationen zu unterschiedlichen Themenfeldern ein. Die Spanne reicht von Pharmakovigilanzanalysen über Behandlungsbewertungen, Risiko-Nutzen-Abwägungen bis hin zu pharmakoökonomischen Fragestellungen. Nach eigenen Angaben führten die Ergebnisse zu Verbesserungen in Bezug auf Arzneimittelsicherheit, Impfempfehlungen und krankheitsspezifischen Behandlungsleitlinien.[167]

165 www.gov.uk/government/news/launch-of-the-clinical-practice-research-datalink (13.12.2021)
166 www.cprd.com/pricing (13.12.2021)
167 https://cprd.com/bibliography (13.12.2021)

5.5.3.5 Einschätzung

Es steht außer Frage, dass die kontinuierliche Sicherheitsüberwachung von Arzneimitteln unter Anwendungsbedingungen eine notwendige Aufgabe im öffentlichen Interesse ist. Weitgehende Übereinstimmung besteht, dass mit den etablierten Verfahren, die sich wesentlich auf klinische Studien und Spontanmeldungen stützen, sehr seltene, mit erheblicher zeitlicher Verzögerung oder in Kombination mit anderen Substanzen auftretende unerwünschte Arzneimittelwirkungen nicht immer zuverlässig erkannt werden können und es aufwendig ist, diesbezügliche Erkenntnisse mit der nötigen Evidenz abzusichern und entsprechendes Wissen zu erweitern. Der retrospektiven Analyse von Daten, die durch die kontinuierliche Behandlungsdokumentation entstehen, wird das Potenzial unterstellt, bei UAW-Hinweisen zeitnah zusätzliche Informationen in Bezug auf damit einhergehende gesundheitsbezogene Risiken generieren zu können (*Spezifikation der Data-Mining-Aufgabe*).

Es steht ebenso außer Frage, dass die Qualität der für derartige Analysen nutzbaren Daten hochrelevant ist. Bei der Prüfung auf unerwünschte Arzneimittelwirkungen betrifft das die zeit- und personenbezogenen Daten zur Arzneimitteleinnahme sowie zu gesundheitlichen Einschränkungen (*Datenauswahl und Aufbereitung*). Im britischen Ansatz kann für diese Aufgabe eine langjährige große Stichprobe medizinischer Behandlungsdaten von Hausärzten verwendet werden, die diese Daten kontinuierlich standardisiert an eine zentrale Forschungsdatenbank übermitteln. Diese Daten werden vielen Forschungsakteuren zugänglich gemacht. Im US-amerikanischen Ansatz wird auf die Datenzusammenführung verzichtet, stattdessen werden aufgabenspezifische Algorithmen an unterschiedliche datenhaltende Stellen geschickt, die ihre Datenbestände damit analysieren und nur Ergebnisse zurückschicken. Da deren medizinische Daten nur begrenzt standardisiert sind, werden administrative (Leistungsabrechnungs-)Daten für deren Pharmakovigilanzuntersuchungen verwendet. In beiden Verfahren können die jeweiligen Daten zeitnah zur Entstehung analysiert werden. Im deutschen Ansatz wurde mit einer großen Stichprobe von Leistungsabrechnungsdaten begonnen (GePaRD), inzwischen werden nahezu alle Leistungsabrechnungsdaten als Vollerhebung über mehrere Etappen geprüft, gespeichert und weitergeleitet (nationaler Versorgungsdatenbestand). Es vergehen Jahre, bis sie in Forschungsdatenbanken eingespeist sind und – bisher eng begrenzt – auf Antrag zugänglich gemacht werden. Die Nutzungsintensität dieser Datenbestände lässt sich von außen kaum beurteilen. Inwiefern die GePaRD-Forschungsdatenbank über die Forschungsdateninfrastruktur perspektivisch intensiver genutzt wird, bleibt abzuwarten. Der GePaRD-Datenbestand deckt zwar einen längeren Zeitraum ab, ist inzwischen jedoch eine Teilmenge des nationalen Versorgungsdatenbestandes. Inwiefern derartige Parallelaktivitäten perspektivisch sinnvoll sind, sollte diskutiert werden. Da die GePaRD-

Datenerhebung früher begann, könnte auch eine Migration der frühen Jahresdatensätze in die nationalen Versorgungsdaten in Erwägung gezogen werden. Schepers et al. (2015, S. 94 f.) sehen weitere Verbesserungsoptionen in der Ausweitung der Medikationserfassung (durch die ambulanten Rezeptverordnungen werden nur gut die Hälfte der in Deutschland abgegebenen Arzneimittel erfasst). Zum einen könnte perspektivisch diskutiert und geprüft werden, ob eine Schnittstelle zu den in Entwicklung befindlichen Medikationsplänen oder sogar die Einbeziehung von patientengeführten Dokumentationen oder digitalen Services, die bei der Arzneimitteleinnahme assistieren, möglich und zulässig seien könnte oder sollte. Zum anderen könnte in Erwägung gezogen werden, Daten zur Medikation in stationären Einrichtungen und Rettungsdiensten zu erfassen. Derartige Überlegungen gehen über die bisherigen Verfahren in erheblichem Maße hinaus. Zum einen sind Arzneimittelabgaben in stationären Einrichtungen bisher kein Bestandteil der Leistungsabrechnung, zum anderen werden Rettungsdienste und stationäre Reha- oder Pflegeeinrichtungen nicht über die gesetzliche Krankenversicherung und das diesbezüglich relevante SGB V reguliert. Mindestens genauso wichtig erscheint dem TAB, die zeitliche Verzögerung zwischen Datengenese und -bereitstellung massiv zu verkürzen.

Die mathematisch-statistischen Verfahren zur Quantifizierung von gesundheitsbezogenen Risiken durch medizinische Maßnahmen sind methodisch umfangreich diskutiert und fachlich allgemein anerkannt (*Datenanalytik, Data-Mining im engeren Sinn*). Klassische statistische Verfahren, die gesundheitsbezogene Risiken anhand komplexer Indexzahlen berechnen, können durch neuere Verfahren ergänzt werden, die auf wahrscheinlichkeitstheoretischen Ansätzen zur Risikoquantifizierung aufbauen.

Die Bewertung der Ergebnisse kann aus unterschiedlichen Perspektiven erfolgen und wird teilweise kontrovers diskutiert (*externe Prüfung, Wissenserweiterung*). Grundsätzlich eignet sich der vorgestellte Ansatz (Disproportionalitätsanalysen) zur Prüfung von Hinweisen auf vorrangig schwerwiegende unerwünschte Ereignisse (sie müssen diagnoserelevant sein), nicht zur Generierung solcher Hinweise (dafür sind klinische Studien, medizinische Forschung und Beobachtung sowie UAW-Melde-Register besser geeignet). Ein (Zusatz-)Nutzen zu den bereits etablierten Pharmakovigilanzelementen entstand im Vioxx©-Fall, als die Vermutung eines erhöhten Gesundheitsrisikos durch eine retrospektive Analyse zeitnah entscheidend fundiert wurde. Der Beleg ist noch nicht erbracht, dass die in Deutschland gestarteten Initiativen, einen substanziellen zusätzlichen Beitrag zu den etablierten Pharmakovigilanzverfahren leisten. Ein solcher Nachweis dürfte schwer zu erbringen sein, denn dafür gibt es kaum Bewertungsmaßstäbe. Zu bedenken ist dabei, dass das arzneimittelrechtlich definierte Qualitätsmanagementsystem bereits sehr hohe Standards bezüglich der Sicherheitsprüfung während der Produktentwicklung und der Sicherheitsüberwachung während der Anwendung setzt, und auch sehr seltene UAW immer

seltener unentdeckt bleiben (dass die US-amerikanischen Datenanalysen nur in wenigen Fällen zu Änderungen der Anwendungsinformationen führten, spricht eher für die etablierten Monitoringsysteme, nicht aber gegen die zusätzlich eingeführten Kontrollmechanismen). Aus dieser Perspektive kann die retrospektive Analyse von Versorgungsdaten als ein Ad-on-Verfahren aufgefasst werden, das wissenschaftlichen Akteuren und Aufsichtsbehörden eigenständige Informationen liefert, die nicht aus herstellerfinanzierten Studien stammen, und damit die Evidenz des Wissens zum Wirkungsspektrum von Arzneimitteln erhöht. Wenn man jedoch nur auf alten Daten rechnen kann, lassen sich UAW kaum zeitnah erkennen. Der britische und der US-amerikanische Ansatz, der statt auf Vollzähligkeit auf Aktualität setzt, scheint diesbezüglich überlegen.

Die methodischen Vorgehensweisen zur Ermittlung von Risikomaßen sind weitgehend akzeptiert, die Algorithmen können immer komplexere Rechenoperationen automatisiert durchführen (*Anwendung zur Entscheidungsunterstützung*). Die Anwendbarkeit und Einsetzbarkeit auf nationalen Leistungsdatenbeständen wurden durch die GePaRD-Aktivitäten belegt. Eine Migration der Verfahren z. B. auf den nationalen Versorgungsdatenbestand sollte möglich sein. Selbst wenn die Rechenschritte automatisiert ausgeführt werden, sind spezifische Kenntnisse erforderlich, um die errechneten Risikowerte einordnen zu können, nächste Schritte abzuwägen sowie Entscheidungen zur Verwendung entsprechender Arzneimittel, Impfstoffe, bis hin zu Behandlungsmethoden oder (digitaler) Medizinprodukte ableiten zu können. Je angreifbarer die zur Analyse verwendeten Daten sind (eingeschränkte Validität aufgrund systemischer Verzerrungen und fehlender Prüfmöglichkeiten), desto schwieriger wird es, Entscheidungen aus den damit ermittelten Ergebnissen im Einzelfall abzuleiten. Die Tragweite derartiger gesundheitssystemischer Entscheidungen ist regelmäßig groß, wie die Fallbeispiele (Kasten 5.1) oder aktuell die Risikobewertung des Coronaimpfstoffs von AstraZeneca zeigen. Das lenkt den Blick noch einmal auf die Daten, die für Vigilanzuntersuchungen im nationalen Gesundheitssystem zugänglich sind. Kritiker/innen unterstellen den Diagnosedaten einen gewissen Abrechnungsbias und dem Verfahren eine begrenzte Prüfbarkeit (Kap. 4.4.1). Vor diesem Hintergrund ist eher erwarten, dass die retrospektive Analyse von Leistungsdaten die etablierten Maßnahmen zur Prüfung der Sicherheit und Verträglichkeit von Arzneimitteln als Add-on-Verfahren ergänzen können, ohne die etablierten Verfahren fundamental zu verändern.

5.6 Fazit

Die informationelle Selbstbestimmung und die ärztliche Schweigepflicht werden im nationalen Gesundheitssystem für vielfältige öffentliche Aufgaben gesetzlich beschränkt. Infolgedessen erhalten unterschiedliche öffentliche Ein-

richtungen des nationalen Gesundheitssystems regelmäßig standardisierte Datensätze mit vielfältigen Bezügen und Informationen: zu Patient/innen und deren gesundheitlicher Situation, zu Ärzt/innen sowie medizinischen Einrichtungen und deren Behandlungsleistungen, zu Krankenkassen und deren Versichertenkollektiven sowie zu Zeit und Raum. Alle Personen und Einrichtungen sind über eindeutige Nummern dauerhaft identifizierbar. Technisch fungieren diese Nummern als Pseudonym und Identifikator, ermöglichen ggf. bei einzelnen datenempfangenden Institutionen zeitliche Bestandsdatenfortschreibungen und Verknüpfungen mit speziellen Registern. Diese maschinell gut verarbeitbaren Versorgungsdaten bilden in der Summe sowohl die gesundheitliche Situation von Versicherten als auch Behandlungs- und Abrechnungsprozesse von medizinischen Einrichtungen und Kostenträgern auf Einzelfallebene im Zeitverlauf vollständig ab. Auch wenn sie keine allzu hohe medizinische Detailgenauigkeit aufweisen, haben sie ein einzigartiges analytisches Potenzial: Versorgungsprozesse, Entwicklungen und gesundheitsbezogene Risiken können hochgranular überwacht sowie gesundheitsbezogene Prozesse verbessert werden. Versorgungsdaten sind personenbeziehbare Daten besonderer Kategorie, bei deren Verarbeitung besondere Schutzmaßnahmen eingehalten werden müssen, um abzusichern, das mit komplexen Analysen und Data-Mining Grundrechte betroffener Personen geschützt werden. Ein komplexes Regelwerk definiert für jede Institution der gesundheitssystemischen Selbstverwaltung

- welche Datenbestände sie aufbauen darf,
- welche Aufgaben sie damit analytisch realisieren soll und
- welche Daten sie ggf. an wen zu welchem Zeitpunkt weiterleiten muss.

Neben diesen gesetzlich definierten primären datenanalytischen Aufgaben (im öffentlichen Interesse) haben einzelne Einrichtungen zudem sekundäre Analysemöglichkeiten, in Bezug auf wissenschaftliche Forschung, Entwicklung und Fortschreibung bestimmter datenanalytischer Ansätze sowie für Planungsaufgaben, bei denen ein Kontrollgremium situativ Schutz- und Nutzungsinteressen abwägt. Für derartige Aktivitäten dürfen Einrichtungen spezielle datenanalytische Abteilungen einrichten oder spezielle Institutionen gründen oder Kooperationen vereinbaren. Auf Anfrage führen einige Institutionen Analysen für Dritte durch– teilweise unentgeltlich, teilweise gegen Gebühr, Apothekenrechenzentren dürfen anonymisierte Rezeptdaten sogar verwerten.

Kassenärztliche Vereinigungen, Apothekenrechenzentren und große gesetzliche Krankenkassen haben eine besondere datenanalytische Position. Sie können große Datenbestände aufbauen, die spezifische Teilbereiche der medizinischen Versorgung zeitnah abbilden, und sie können diese Versorgungsdaten mit hoher Aktualität monopolisiert analysieren. Dritten werden Versorgungsdaten über die Forschungsdatenzentren des DIMDI/BfArM und der statistischen Ämter (teilweise) mit mehrjähriger Verzögerung zugänglich gemacht. Bisher

war das Nutzungsinteresse insbesondere an den vom DIMDI/BfArM bereitgestellten Versorgungsdaten begrenzt. Ein Grund ist die geringe Aktualität der Daten, ein anderer die engen Grenzen der Nutzungsberechtigung. Ausschließlich Institutionen der gesundheitssystemischen Selbstverwaltung und der wissenschaftlichen Forschung sind antragsberechtigt. Einige von ihnen (z. B. Krankenkassen) können jedoch aktuellere Teildatenbestände unmittelbar nutzen und dafür teilweise mit akademischen Institutionen kooperieren.

Zwar legitimieren die primären Aufgaben im öffentlichen Interesse die Beschränkungen sowohl der informationellen Selbstbestimmung als auch der ärztlichen Schweigepflicht. Das Fehlen jeglicher Widerspruchsmöglichkeiten für Betroffene in sekundäre Weiterverwendungen derartiger Daten besonderer Kategorie wird mitunter jedoch als paternalistische Fremdbestimmung kritisiert, zumal unterschiedliche Teilbestände in mehreren Etappen weitergeleitet und von unterschiedlichen Institutionen mit immer neuen Aufgaben und Weiterverwendungsmöglichkeiten verwendet werden können. Die generelle Einschätzung, dass betroffene Personen keine Kontrolle über Daten mit Bezügen zu ihrer Person haben, es den verketteten Datenweiterleitungen und Verarbeitungsmöglichkeiten unterschiedlicher Institutionen an Transparenz mangele und betroffene das Vorgehen kaum überblicken können, dürfte auch auf die datenverarbeitenden Prozesse des nationalen Gesundheitssystems zutreffen. Auch der Begriff der Datentransparenz wird im nationalen Gesundheitssystem eigenständig interpretiert. Eine öffentlich zugängliche Darstellung von Nutzungsabsichten, ggf. durchgeführter Data-Mining-Prozesse oder gewonnener Erkenntnisse wird bisher nicht damit assoziiert.

Gesundheitssystemische Data-Mining-Prozesse starten in der Regel zunächst in kleinerem Rahmen als Forschungsprojekte oder als Machbarkeitsstudien, deren Ergebnisse anschließend fachlich diskutiert werden. Dabei wird regelmäßig deutlich, dass auch methodisch und analytisch geeignete Verfahren nur solche Strukturen und Informationen extrahieren können, die in den Analysedatensätze enthalten sind. Eine räumlich zu geringe Auflösung kann keine lokalen Spezifika aufzeigen, alte Analysedaten können keine Risiken neuer Arzneimittel oder Behandlungsmethoden zeitnah sichtbar machen. Die jeweiligen datenanalytischen Ansätze von Data-Mining-Prozessen und die resultierenden Ergebnisse werden in der Regel in Fachkreisen diskutiert, situativ abgewogen und bewertet. Danach können Verfahren ggf. verstetigt und Algorithmen z. B. in epidemiologische Informationsdienste oder in größere gesundheitssystemische Prozesse, wie das Fallpauschalensystem, integriert werden. Einen Produktstatus erreichen dieserart Algorithmen in der Regel nicht.

6 Gesamtfazit und Handlungsoptionen

6.1 Fazit

Data-Mining ist ein unscharfer Begriff – ähnlich wie die Big oder Smart Data, maschinelles Lernen oder künstliche Intelligenz. Es zeigen sich hier erhebliche Schnittmengen, insbesondere wenn man die damit einhergehenden gesellschaftlichen Herausforderungen in den Blick nimmt. Denn alle Begriffe werden mit datenanalytischen Verfahren assoziiert, die aus großen (Trainings-)Datenbeständen Strukturen extrahieren, Regeln ableiten oder Modelle anpassen. Data-Mining wird tendenziell eher mit wissenschaftlichen Forschungsaktivitäten in Verbindung gebracht als die anderen Begriffe, weil Data-Mining mit dem Ziel assoziiert wird, Informationen oder Erkenntnisse zu Datenstrukturen zu gewinnen, diesbezügliches Wissen zu generieren bzw. zu erweitern. Mit diesen Begriffen verbundene Visionen beruhen oftmals auf der Grundannahme, dass immer mehr Daten auch komplexe reale Phänomene so umfangreich und genau abbilden, dass Regeln und Modelle weitgehende Allgemeingültigkeit erreichen und zur Klassifikation und Gruppierung neuer Sachverhalte oder zur Prognose von Entwicklungen eingesetzt werden können. Vielfältige Praxisbeispiele von genetischen Tests über die Streckenoptimierung von Navigationssystemen bis zu Klimamodellen stützen diese Annahme. Zugleich betonen Datenexpert/innen, dass zum einen auch große Datenmengen reale Sachverhalte in ihrer Vielschichtigkeit kaum vollumfänglich abbilden und Regeln und Modelle stets Vereinfachungen einer komplexeren Realität seien. Zum anderen weisen sie darauf hin, dass durch derartige datenbasierte Vorgehensweisen real existierende strukturelle Probleme, wie z. B. die Diskriminierung einzelner Personengruppen reproduziert werden könnten. Auch dafür gibt es Belege aus der Praxis, z. B., dass Algorithmen höhere Rückfallwahrscheinlichkeiten bei afroamerikanischen Straftäter/innen ermittelten oder Männer in Bewerbungssituationen als geeigneter einstuften. Folglich können die Resultate derartiger Prozesse in konkreten einzelnen Anwendungskontexten nützlich sein, aber auch Risiken allgemeiner oder auch ganz neuer Art mit sich bringen. Die Schaffung eines Mehrwerts unter Achtung der freiheitlichen Grundordnung ist folglich eine Frage der sinnvollen Ausgestaltung derartiger Prozesse.

6.2 Allgemeine Handlungsoptionen

Zahlreiche Sachverständigenräte und Kommissionen auch des Deutschen Bundestages und der Bundesregierung haben sich in den letzten Jahren mit den

Möglichkeiten und Herausforderungen der Digitalisierung im Allgemeinen sowie den wachsenden Datenbeständen, mit den Möglichkeiten und Grenzen deren Analyse und mit dem Umgang der Ergebnisse im Besonderen auseinandergesetzt sowie diesbezüglich Empfehlungen und Handlungsoptionen erarbeitet, zu denen wiederum zahlreiche Stakeholder Stellung genommen haben. Unisono wird empfohlen, Digitalisierungsaktivitäten zu forcieren, Infrastrukturen zur Weiterverwendung von Datenbeständen auf- und auszubauen, die Datennutzung stärker in den Blick zu nehmen, entsprechendes Know-how zu stärken, die Entwicklung datenanalytischer Anwendungen zu fördern, risikoreiche Anwendungen stärker zu regulieren sowie eine größere nationale oder europäische digitale Souveränität anzustreben, auch um hohe Schutzstandards und die Grundrechtesicherung zu gewährleisten. Diese Empfehlungen lassen sich auch aus den Ausführungen dieses Berichts und den dafür in Auftrag gegebenen Gutachten ableiten.

6.3 Handlungsoptionen, die sich aus dem Vergleich der Fallbeispiele ableiten lassen

Bei einer vergleichenden Betrachtung unterschiedlicher datenanalytischer Anwendungsbereiche wird zudem deutlich, dass es bereichsspezifische Besonderheiten, Stärken und Schwerpunktsetzungen gibt, die sich teilweise zu ergänzen scheinen. Eine abschließende vergleichende Gesamtschau soll Handlungsoptionen für das Parlament fundieren.

Standardisierung, Zugänglichkeit und Nutzbarkeit von Daten verbessern

Interoperable Datenzugangsstrukturen sind notwendige Bedingungen für Data-Mining. Der Geodatenbereich mit seinen Gremien und langjährigen Aktivitäten zum Aufbau der nationalen Dateninfrastruktur gilt diesbezüglich als ein Vorreiter. Hier haben sich Normen und Standards bei der Datenerfassung, -speicherung und -analyse weitgehend durchgesetzt. Über die nationale Geodateninfrastruktur können standardisierte amtliche Referenzdatenbestände bereitgestellt werden. Die ursprüngliche Differenzierung der Datenbereitstellung für öffentliche Aufgaben, Forschungstätigkeiten und kommerzielle Weiterverwendungsabsichten wird zunehmend aufgegeben, Open-Data-Ansätze gewinnen an Bedeutung und vormalige Zugangshürden sinken. Das datenanalytische Potenzial dieser Geobasisdaten steigt, je mehr Fachdaten aus anderen Bereichen auf kommunaler, Landes- und Bundesebene georeferenziert und mit ihnen verknüpft werden. Dazu sind erhebliche Aktivitäten auch in anderen Fachbereichen nötig.

Zwar werden öffentliche Einrichtungen zunehmend zur Bereitstellung ihrer georeferenzierten Daten verpflichtet, inwiefern diese aber für komplexe Datenanalysen tatsächlich genutzt werden, lässt sich bisher nur schwer abschätzen. Antrags- oder Nutzungsregister, die Weiterverwendungsabsichten, datennutzende Projekte oder die Entwicklung von spezifischen Informationsprodukten und -diensten skizzieren, gibt es bisher kaum. Die dem Deutschen Bundestag am Ende jeder Legislaturperiode vorzulegenden Geo-Fortschrittsberichte thematisierten bisher vor allem Aktivitäten zur Datenbereitstellung, die Nachfrageentwicklung und die Datenweiterverwendung hingegen weniger. Zukünftig sollten die Datennachfragen und Datenweiterverwendungen stärker in den Blick genommen werden, um die Potenziale der Datenangebote gezielter erfassen, bewerten und auszuschöpfen zu können. Der Deutsche Bundestag könnte diesbezügliche Untersuchungen in den regelmäßig vorzulegenden Fortschrittsberichten einfordern.

Einrichtungen des Gesundheitssystems wird seit Jahren erheblicher Entwicklungsbedarf bezüglich der Digitalisierung unterschiedlicher datenverarbeitender Prozesse, der Entwicklung und Nutzung von Datenstandards und des Aufbaus interoperabler Datenzugangsstrukturen attestiert. Ärzt/innen sind zwar zur Erhebung medizinisch notwendiger Daten und zur Behandlungsdokumentation in arztgeführten Primärakten verpflichtet, bisher gibt es jedoch keine verbindlichen Vorgaben zur Verwendung einheitlicher Terminologien und interoperabler (Datei-)Formate. Um den zukünftigen Aufwand für die analysevorbereitenden Datenaufbereitungen zu senken, sollte die Entwicklung und Verwendung einheitlicher medizinischer Terminologien und interoperabler Formate bereits bei der primären Behandlungsdokumentation vorangetrieben und perspektivisch vorgeschrieben werden. Dabei gilt es den Arbeitsaufwand von behandelnden Ärzt/innen im Blick zu behalten und nach Lösungen zu suchen, die den Dokumentationsaufwand so gering wie möglich halten.

Die Daten der arztgeführten Primärakten unterliegen der Schweigepflicht und höchsten Datenschutzvorgaben. Sie werden in spezifischen Informationssystemen einrichtungsintern gespeichert und archiviert. Diese Systeme sind nicht für medizinische Data-Mining-Aktivitäten konzipiert. Um Behandlungsdaten dafür weiterzuverwenden, müssen diese aufbereitet und in zumeist einrichtungsübergreifende sekundäre Register, Repositorien oder Datenzentren überführt werden. Dafür sind gesetzliche Regelungen (bei Aufgaben im öffentlichen Interesse) oder Einwilligungen erforderlich, die bisher schriftlich eingeholt werden. Beide Verfahrensformen werden seit langem genutzt, um diverse, spezifisch definierte Datensätze aus den Primärakten abzuleiten und an unterschiedliche medizinische Register oder Datenzentren zu übermitteln, die diese Daten für administrative und gesundheitssystemische Aufgaben, aber auch zu Forschungs- und Planungszwecken bereitstellen. Diese Register und Datenzentren fungieren als Datentreuhänder in vielfältigen Organisationsformen. Diese

bereits etablierten Datentreuhandformen sollten bezüglich ihrer Praktikabilität geprüft, weiterentwickelt und harmonisiert werden. Sie könnten beispielgebend auch für andere Bereiche sein, in denen geschützte Daten nicht monopolisiert gehalten, sondern unter Einhaltung ethischer Standards weiterverwendet werden sollen (z. B. Mobilitätsdaten).

Im Laufe der Zeit haben vielfältige spezialgesetzliche Regelungen zum Umgang mit gesundheitsbezogenen Daten in den unterschiedlichen Einrichtungen des Gesundheitssystems eine erhebliche Komplexität erreicht, die keinesfalls leicht zu erfassen ist, zu Unsicherheiten bezüglich der Möglichkeiten und Grenzen der Datenweiterverwendung führt und dadurch Datenanalysen erschwert. Auch gibt es weder einen genauen Überblick über die Vielfalt der Register und Datenzentren mit ihren jeweiligen Datenbeständen und Nutzungsmöglichkeiten noch eine übergeordnete Dateninfrastruktur, die diese Datenzentren und Register vernetzt und die Daten des nationalen Gesundheitswesens unter Einhaltung der bestehenden datenschutzrechtlichen und medizinethischen Normen zugänglich macht. Das 2021 verabschiedete Datennutzungsgesetz, das darauf abzielt, die Nutzungsmöglichkeiten der Daten des öffentlichen Sektors zu harmonisieren und zu befördern, gilt nicht für die Daten, die im nationalen Gesundheitssystem verarbeitet werden. Ein diesbezügliches Spezialgesetz, das die Vielfalt der gesundheitssystemischen datenbezogenen Regelungen harmonisiert und vereinfacht, erscheint daher dringend geboten.

Mit der seit 2021 allen Versicherten anzubietenden elektronischen Patientenakte werden derzeit große Hoffnungen verbunden, vielfältige gesundheitsbezogene Daten vor allem aus Behandlungskontexten in der Verantwortung einzelner Patient/innen zusammenzuführen und perspektivisch auch das Einwilligungsmanagement für die Datenweitergabe bis hin zu Datenspenden zu Forschungszwecken damit zu organisieren. Dieses Einwilligungsmanagement ist von zentraler Bedeutung für Datenweiterverwendungen einschließlich Data-Mining. Bisher werden Einwilligungen schriftlich eingeholt. Mit der elektronischen Patientenakte könnten Versicherte ihre Einwilligungen perspektivisch digital organisieren. Wie viele Versicherte dieses Angebot zur Datenselbstverwaltung annehmen und in die Datenweiterverwendung zu Forschungszwecken einwilligen werden, ist derzeit noch unklar. Eine Begleitforschung zur Entwicklung zur Akzeptanz dieser Akten und der Nutzung unterschiedlicher Funktionalitäten und Services scheint dringend geboten. Das Parlament könnte sich berichten lassen.

Konkretion der privilegierten Datenverwendung zu Forschungszwecken

Datenweiterverwendungen zu wissenschaftlichen Forschungszwecken einschließlich Data-Mining werden zum einen datenschutzrechtlich privilegiert,

zum anderen begrenzen sie Urheber- bzw. Leistungsschutzrechte. Etliche Formulierungen zum Forschungsprivileg sind jedoch auslegungswürdig. Die DSGVO empfiehlt lediglich, den wissenschaftlichen Forschungsbegriff mit der Einhaltung anerkannter ethischer Forschungsstandards zu verknüpfen, ein entsprechendes Einwilligungsmanagement vorzusehen, Forschungsabsichten im Einzelfall zu prüfen und sowohl öffentliche als auch privatwirtschaftlich finanzierte Forschung bis hin zu technologischen Entwicklungen und Anwendungsdemonstrationen zuzulassen. Über Öffnungsklauseln lässt sie jedoch nationale Spezifikationen zu.

Ethische (Forschungs-)Standards und die Prüfung von Datennutzungsanträgen sind in der Medizin und im Gesundheitssystem seit langem in besonderem Maße verankert. Die Auseinandersetzung mit (medizin-)ethischen Prinzipien von der Schadensvermeidung über die Einhaltung wissenschaftlicher Gütekriterien bis zur Achtung der informierten Selbstbestimmung und der Einhaltung hoher Datenschutzanforderungen beginnt bereits in der medizinischen Ausbildung. Diese Prinzipien sind zudem im Berufsrecht verankert und sie bilden den normativen Rahmen für die Prüfung von Forschungsanträgen einschließlich sekundärer Datenverwendungen und Data-Mining. Handlungsbedarf gibt es derzeit vor allem bezüglich der Vereinheitlichung, Beschleunigung und Straffung der Antragsprüfungen. Bezüglich der ethischen Standards und der Prüfung von Datennutzungsanträgen ist das Gesundheitssystem ein Vorreiter, von dem andere Bereiche lernen könnten. Das Einwilligungsmanagement in Datenweiterverwendungen ist derzeit bei medizinischen und gesundheitssystemischen Einrichtungen jedoch eine gewisse Schwachstelle. Bisher können Einwilligungen und Widerrufe nur schriftlich erteilt werden, was vor allem bei umfangreichen Datenweiterverwendungen rückwirkend kaum machbar ist. Auch aus diesem Grund wird der Forschungsbegriff abweichend von der DSGVO im deutschen Gesundheitssystem eng ausgelegt. Für die in unterschiedlichen Registern und Datenzentren gespeicherten personenbezogenen Gesundheitsdaten sind in der Regel nur bestimmte öffentliche (Forschungs-)Einrichtungen nutzungsberechtigt, Forschungsabsichten müssen im öffentlichen Interesse liegen. Dadurch können u. a. Medizinproduktehersteller nur in Kooperation mit öffentlichen Forschungseinrichtungen entsprechende Daten nutzen, um z. B. algorithmische Assistenzsysteme zu trainieren.

Parallel dazu sind Unternehmen, die klinische Studien finanzieren, mit denen die Sicherheit und Wirksamkeit bzw. Leistungsfähigkeit von diagnostischen oder therapeutischen Verfahren geprüft werden, nicht dazu verpflichtet, ihre Studiendaten Dritten zugänglich zu machen.

Vertreter/innen der freien Wirtschaft, der (medizinischen) Forschung sowie öffentlicher Einrichtungen kritisieren seit Jahren die derzeitigen Verfahren sowie die damit einhergehenden Ungleichbehandlungen und betonen im medizinischen Kontext die gesundheitsbezogenen Risiken durch die Nichtnutzung von Daten, wenn beispielsweise Erkrankungsrisiken, Infektionsherde oder unerwünschte

Nebenwirkungen von Behandlungsverfahren nicht erkannt werden. Die Etablierung offenerer Datennutzungskonzepte sollte daher diskutiert bzw. geprüft werden. Dazu könnten die Reichweite des Forschungsbegriffs und bestehende Datenverarbeitungsprivilegien diskutiert und gesetzlich klargestellt werden.

Qualitätsmanagementsysteme bei Medizinprodukten – Vorbild für den Umgang mit Data-Mining-Ergebnissen in anderen Bereichen?

Inwiefern Data-Mining-Prozesse das Gemeinwohl steigern, dabei die Grundrechte Einzelner schützen oder gefährden, transparent gestaltet sind oder aber mit menschlichen Kontrollverlusten in Entscheidungsprozessen einhergehen und welche Folgen daraus erwachsen, kann nur situativ abgewogen und bewertet werden. Die im medizinischen Kontext über Jahrzehnte entstandenen Verfahren zur Qualitätssicherung medizinischer Produkte mit ihren risikoajustierten abgestuften Zertifizierungsverfahren in Kombination mit kontinuierlichen produktbezogenen Sicherheitsprüfungen und Risikoüberwachungen während der Anwendung könnten beispielgebend für andere risikoreiche Anwendungsbereiche sein, in denen datenanalytische Verfahren und algorithmenbasierte Systeme zunehmend eingesetzt werden (z. B. innere oder äußere Sicherheits-, Fin- oder Legal-Tech-Bereiche). Die Forderungen nach risikoadaptierten Regulierungen und Algorithmen-TÜVs oder der derzeit auf europäischer Ebene verhandelte Digital Service Act greifen unterschiedliche qualitätssichernde Maßnahmen des Medizinprodukterechts bereits auf. Mit diesbezüglichen Vorgehensweisen, deren Konkretisierung und Harmonisierung vor allem in risikoreichen Anwendungskontexten sollten Stakeholder sich intensiver befassen. Dadurch könnten Analyst/innen und Prüfinstanzen wichtige Informationen zur Sicherheits- und Leistungsbewertung erhalten sowie Prüf- und Monitoringverfahren etabliert werden, mit denen Risiken während der Anwendung algorithmischer Systeme überwacht und ggf. reduziert werden könnten.

Die unterschiedlichen Elemente der im medizinischen Bereich etablierten Qualitätsmanagementsysteme zielen primär auf eine hohe Produktsicherheit und die Generierung eines gesundheitsbezogenen Nutzens durch die Produktanwendung ab. Jedoch lassen sich auch mit höchst umfangreichen Qualitätsmanagementsystemen beim Einsatz datentrainierter algorithmischer Systeme im Rahmen der Behandlung nie alle Risiken für Betroffene vollständig ausschließen, denn auch große Datenbestände und komplexe mathematisch-statistische Modelle bilden die Realität vereinfacht ab, kommen bei höchst seltenen Situationen an ihre Grenzen, können real existierende Diskriminierungen reproduzieren und liefern Ergebnisse, die mitunter selbst für Expert/innen im Detail nur schwer nachzuvollziehen sind. Deshalb sind die Klärung von dauerhaften Produktverantwortlichkeiten und von Haftungsfragen relevante Aspekte für die Akzeptanz und den Einsatz algorithmischer Assistenzsysteme. Forschungseinrichtungen, die Daten privilegiert nutzen dürfen, um Modelle zu trainieren und

Assistenzsysteme zu entwickeln, kommen regelmäßig bereits bei der Produktzertifizierung an ihre Grenzen. Die kontinuierliche Gewährleistung einer hohen Produktsicherheit und Haftung im Schadenfall gehört nicht mehr in das Tätigkeitsspektrum von Forschungseinrichtungen. Spätestens dafür sind wirtschaftlich agierende Unternehmen erforderlich. Bereits bei klassischen Softwareprodukten wird die Eignung des derzeitigen Haftungsrechts in medizinischen, aber auch in anderen Einsatzbereichen kontrovers diskutiert. Besondere haftungsrechtliche Herausforderungen ergeben sich durch kontinuierlich lernende, medizinische Assistenzsysteme. Verantwortlichkeiten und Haftungsfragen bis hin zu Härtefallfonds zum Schadensausgleich sollten daher systematisch und spezifisch durchdacht, abgewogen und rechtlich geklärt werden.

7 Literatur

7.1 In Auftrag gegebene Gutachten

Bernsdorf, B.; Bierbrauer, H.; Büscher, O.; Müterthies, A.; Pakzad, K.; Wenzel, T.; Woditsch, S. (2015): Data-Mining: Gesellschaftspolitische und rechtliche Herausforderungen. Data-Mining mit Geodaten (Fallstudie 2). Münster

Schepers, J.; Schlünder, I.; Drepper, J.; Semler, S.; Rüping, S.; Quix, C.; Stroetmann, K.; Rennoch, J. (2015): Data-Mining in der Medizin und im Gesundheitssystem − gesellschaftspolitische und rechtliche Herausforderungen. Berlin

7.2 Weitere Literatur

Ada (2018): Ada startet Global Health Initiative. Presseerklärung, https://assets.ctfassets.net/jsvgavb9trbp/3JIa17qCuA0wUgsuIwqYKc/f3fe8b286c27c0e4702735c8d01aed79/181009_Pressrelease-GHI_DE.pdf (13.12.2021)

AG NGIS (Arbeitsgruppe NGIS des Lenkungsgremium GDI-DE) (2015): Nationale Geoinformations-Strategie. https://www.gdi-de.org/download/NGIS_Nationale_Geoinformationsstrategie_V1.pdf (13.12.2021)

Akademien der Wissenschaften Schweiz (2015): Big Data im Gesundheitswesen. www.samw.ch/dam/jcr:93263052-6f12-4ab8-bcfa-821b640fe225/white_paper_samw_big_data_gesundheitswesen.pdf (13.12.2021)

Angerer, C. (2018): Neuronale Netze. Revolution für die Wissenschaft? In: Spektrum der Wissenschaft (1), S. 12–19

Antes, G.; Bertram, I. (2019): Big Data, big Errors. In: Gen-ethischer Informationsdienst (248), S. 10–11

ARGE IK (Arbeitsgemeinschaft Institutionskennzeichen) (2015): Gemeinsames Rundschreiben Institutionskennzeichen (IK). www.gkv-datenaustausch.de/media/dokumente/leistungserbringer_1/Gemeinsames_Rundschreiben_IK_2015-03.pdf (13.12.2021)

Baas, J.; Möws, V. (2017): »Jede Ergänzung des RSA (sollte) sicher vor Manipulationen sein«: Zum Kodierwettbewerb der Krankenkassen. In: RPG 23(1), S. 3–9

Baas, J.; Scherff, D. (2016): »Wir Krankenkassen schummeln ständig«. Frankfurter Allgemeine Sonntagszeitung, www.faz.net/aktuell/finanzen/meine-finanzen/versichern-und-schuetzen/interview-mit-jens-baas-chef-der-techniker-krankenkasse-14472241.html (13.12.2021)

BAG Selbsthilfe (2020): Stellungnahme. Referentenentwurf. Referentenentwurf einer Verordnung zur Neufassung der Datentransparenzverordnung und zur Änderung der Datentransparenz-Gebührenverordnung. www.bag-selbsthilfe.de/fileadmin/user_upload/News/2020/Stellungnahme_zum_Referentenentwurf_einer_Verordnung_zur_Neufassung_der_Datentransparenzverordnung_und_zur_AEnderung_der_Datentransparenz-Gebuehrenverordnung.docx (13.12.2021)

Balling, S. (2018): Medizin nach dem Gusto der Kassen? In: f&w 10, S. 876–870

Balzter, S. (2018): Supercomputer Watson. Im Krankenhaus fällt die Wunderwaffe durch. Frankfurter Allgemeine Zeitung, www.faz.net/aktuell/wirtschaft/kuenstli

che-intelligenz/computer-watson-scheitert-zu-oft-bei-datenanalyse-15619989/das-computersystem-watson-soll-15620798.html (13.12.2021)
Bayrische Landesapothekerkammer (2017): Berufsordnung für Apothekerinnen und Apotheker. www.blak.de/berufsordnung (13.12.2021)
Beauchamp, T.; Childress, J. (2008): Principles of Biomedical Ethics. Oxford
Becker, A.; Marcon, M.; Ghafoor, S.; Wurnig, M.; Frauenfelder, T.; Boss, A. (2017): Deep learning in mammography: Diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. In: Investigative Radiology 52(7), S. 434–440
Beerheide, R. (2019): Genexpressionstest: Debatte um Datenauswertung. In: Deutsches Ärzteblatt 116, S. 29–30
Behrends, S. (2018): Mustergültige Diagnosen – Wie KI die Medizin verändert. Interview mit Forsting, M. In: Chefsache Patientenversorgung 1, S. 16–19
Beivers, A.; Emde, A. (2020): DRG-Einführung in Deutschland: Anspruch, Wirklichkeit und Anpassungsbedarf aus gesundheitsökonomischer Sicht. In: Klauber, J.; Geraedts, M.; Friedrich, J.; Wasem, J.; Beivers, A. (Hg.): Krankenhaus-Report 2020. Finanzierung und Vergütung am Scheideweg. Berlin, S. 5–24
BfArM (Bundesinstitut für Arzneimittel und Medizinprodukte) (2016): Neben-wirkungen von Arzneimitteln melden: Europaweite Kampagne soll Patientinnen und Patienten sensibilisieren. Pressemitteilung Nr. 11, www.bfarm.de/SharedDocs/Pressemitteilungen/DE/2016/pm11-2016.html (17.12.2021)
BfDI (Bundesbeauftragter für den Datenschutz und die Informationsfreiheit) (2020): Tätigkeitsbericht für die Jahre 2018 und 2019 zur Informationsfreiheit. Deutscher Bundestag, Drucksache 19/19910, Berlin
BfR (Bundesinstitut für Risikobewertung) (2014): Fragen und Antworten zu Methicillin-resistenten Staphylococcus aureus (MRSA) – Aktualisierte FAQ vom 18.11.2014. Berlin, www.bfr.bund.de/cm/343/fragen-und-antworten-zu-methicillin-resistenten-staphylococcus-aureus-mrsa.pdf (13.12.2021)
BGH (Bundesgerichtshof): Die »DWD WarnWetter-App« darf nur für Wetterwarnungen kostenlos und werbefrei angeboten werden. Pressemitteilung 28, www.bundesgerichtshof.de/SharedDocs/Pressemitteilungen/DE/2020/2020028.html (13.12.2021)
Bill, R.; Fritsch, D. (1991): Grundlagen der Geo-Informationssysteme. Band 1: Hardware, Software und Daten. Karlsruhe
BIPS (Leibniz-Institut für Präventionsforschung und Epidemiologie) (2017): German Pharmacoepidemiological Research Database. www.bips-institut.de/fileadmin/bips/images/gepard/GePaRD_description_V1.9.pdf (13.12.2021)
Bishop, C. M. (2006): Pattern Recognition and Machine Learning. New York
Bitkom (2015): Kognitive Maschinen – Meilenstein in der Wissensarbeit. www.bitkom.org/sites/default/files/file/import/150213-Kognitive-Maschinen-11Febr 2015.pdf (13.12.2021)
Bitkom (2019): Blick in die Blackbox. Nachvollziehbarkeit von KI-Algorithmen in der Praxis. www.bitkom.org/sites/default/files/2019-10/20191016_blick-in-die-blackbox.pdf (13.12.2021)
BMBF (Bundesministerium für Bildung und Forschung) (2020): BMBF-Aktionsplan Forschungsdaten. www.bmbf.de/files/163_20_Faktenblatt_Aktionsplan_4.pdf (13.12.2021)
BMWi (Bundesministerium für Wirtschaft und Energie) (2018): Monitoring-Report. Wirtschaft digital 2018. www.bmwi.de/Redaktion/DE/Publikationen/Digitale-

Welt/monitoring-report-wirtschaft-digital-2018-kurzfassung.pdf?__blob=publicationFile&v=24 (13.12.2021)
BMWi (2019): Das Projekt GAIA-X. Eine vernetzte Dateninfrastruktur als Wiege eines vitalen, europäischen Ökosystems. www.bmwi.de/Redaktion/DE/Publikationen/Digitale-Welt/das-projekt-gaia-x.pdf?__blob=publicationFile&v=24 (13.12.2021)
Bodammer, L.; Scholz, A.; Engelbrecht, S. Kandler, A. (2014): Keine Nutzung von Patientendaten durch Marktforschungsunternehmen ohne persönliches Einverständnis. Entschließungsantrag. https://www.bundesaerztekammer.de/arzt2014/media/applications/EVII82.pdf (13.12.2021)
Bond, W.; Schwartz, L.; Weaver, K.; Levick, D.; Giuliano, M.; Graber, M. (2012): Differential diagnosis generators: an evaluation of currently available computer programs. In: Journal of general internal medicine 27(2), S. 213–219
Borchardt, F. (2012): Krankenhaus-Rechnungsprüfungen. Spannungsfeld zwischen Konflikt und Kooperation. Verband der Ersatzkassen, www.vdek.com/magazin/ausgaben/2012-0708/titel-krankenhaus-rechnungspruefungen.html (13.12.2021)
Brinkman, A.; Nik-Zainal, N.; Simmer, F.; Rodríguez-González, F.; Smid, M.; Alexandrov, L.; Butler, A.; Martin, S.; Davies, H.; Glodzik, D.; Zou, X.; Ramakrishna, M. et al. (2019): Partially methylated domains are hypervariable in breast cancer and fuel widespread CpG island hypermethylation. In: Nature Communications 10, Art. 1749
Brown, J.; Lane, K.; Moore, K.; Platt, R. (2009): Defining and Evaluating Possible Database Models to Implement the FDA Sentinel Initiative. www.brookings.edu/wp-content/uploads/2012/04/03_Brown.pdf (13.12.2021)
BSI (Bundesamt für Sicherheit in der Informationstechnik) (2017): Schutz Kritischer Infrastrukturen. durch IT-Sicherheitsgesetz und UP Kritis. www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Publikationen/Broschueren/Schutz-Kritischer-Infrastrukturen-ITSig-u-UP-KRI-TIS.pdf;jsessionid=BDD80015069CF8DE62A413465ABDD1AA.internet471?__blob=publicationFile&v=1 (13.12.2021)
Bundesrechnungshof (2019): Bericht an den Rechnungsprüfungsausschuss des Haushaltsausschusses des Deutschen Bundestages nach § 88 Abs. 2 BHO. über die Prüfung der Krankenhausabrechnungen durch die Krankenkassen der gesetzlichen Krankenversicherung. www.bundesrechnungshof.de/de/veroeffentlichungen/produkte/beratungsberichte/langfassungen/langfassungen-2019/2019-bericht-krankenhausabrechnungen-durch-die-krankenkassen-der-gesetzlichen-krankenversicherung-pdf/@@download/file (13.12.2021)
Bundesregierung (2003): Geoinformationspolitik in Deutschland. Antwort der Bundesregierung auf die Kleine Anfrage der Abgeordneten Ernst Burgbacher, Daniel Bahr (Münster), Rainer Brüderle, weiterer Abgeordneter und der Fraktion der FDP – Drucksache 15/1157 –. Deutscher Bundestag, Drucksache 15/1209, Berlin
Bundesregierung (2005): Bericht der Bundesregierung über die Fortschritte zur Entwicklung der verschiedenen Felder des Geoinformationswesens im nationalen, europäischen unter internationalen Kontext. Unterrichtung durch die Bundesregierung. Deutscher Bundestag, Drucksache 15/5834, Berlin
Bundesregierung (2008): Zweiter Bericht der Bundesregierung über die Fortschritte zur Entwicklung der verschiedenen Felder des Geoinformationswesens im nationalen, europäischen und internationalen Kontext. Unterrichtung durch die Bundesregierung. Deutscher Bundestag, Drucksache 16/10080, Berlin

Bundesregierung (2012a): Dritter Bericht der Bundesregierung über die Fortschritte zur Entwicklung der verschiedenen Felder des Geoinformationswesens im nationalen, europäischen und internationalen Kontext. Deutscher Bundestag, Drucksache 17/11449, Berlin

Bundesregierung (2012b): Entwurf eines Gesetzes zur Verbesserung der Rechte von Patientinnen und Patienten. Deutscher Bundestag, Drucksache 17/10488, Berlin

Bundesregierung (2016a): Chancen der Nutzung von Open Data. Antwort auf die kleine Anfrage der Fraktion BÜNDNIS 90/DIE GRÜNEN. Deutscher Bundestag, Drucksache 18/7485, Berlin

Bundesregierung (2016b): Entwurf eines Gesetzes zur Änderung des Bundesstatistikgesetzes und anderer Statistikgesetze. Deutscher Bundestag, Drucksache 18/7561, Berlin

Bundesregierung (2016c): Stand der Weiterentwicklung des pauschalierenden Entgeltsystems Psychiatrie und Psychosomatik. Deutscher Bundestag, Deutscher Bundestag, Drucksache 18/7281, Berlin

Bundesregierung (2017): Vierter Bericht der Bundesregierung über die Fortschritte zur Entwicklung der verschiedenen Felder des Geoinformationswesens im nationalen, europäischen und internationalen Kontext (4. Geo-Fortschrittsbericht der Bundesregierung). Unterrichtung durch die Bundesregierung. Deutscher Bundestag, Drucksache 18/12872, Berlin

Bundesregierung (2018a): Big Data, Scoring und Datenhandel von Parteiaffinitäten der Bundesbürger durch die Deutsche Post Direkt GmbH. Antwort auf die Kleine Anfrage der Fraktion BÜNDNIS 90/DIE GRÜNEN. Deutscher Bundestag, Drucksache 19/2150, Berlin

Bundesregierung (2018b): Stand der Einführung des elektronischen Gesundheitsberuferegisters. Antwort auf die Kleine Anfrage der Fraktion BÜNDNIS 90/DIE GRÜNEN. Deutscher Bundestag, Drucksache 19/4185, Berlin

Bundesregierung (2018c): Strategie Künstliche Intelligenz der Bundesregierung. www.bmwi.de/Redaktion/DE/Publikationen/Technologie/strategie-kuenstliche-intelligenz-der-bundesregierung.pdf?__blob=publicationFile&v=10 (13.12.2021)

Bundesregierung (2018d): Zukunft der elektronischen Gesundheitskarte. Antwort auf die Kleine Anfrage der Fraktion der FDP. Deutscher Bundestag, Drucksache 19/2358, Berlin

Bundesregierung (2019a): Bericht der Bundesregierung zur Evaluierung des Gesetzes zur Förderung der elektronischen Verwaltung sowie zur Änderung weiterer Vorschriften. Deutscher Bundestag, Drucksache 19/10310, Berlin

Bundesregierung (2019b): Datenschutz und Beratung im Arzneimittelversandhandel. Antwort auf die Kleine Anfrage der Fraktion DIE LINKE. Deutscher Bundestag, Drucksache 19/7831, Berlin

Bundesregierung (2019c): Eckpunkte einer Datenstrategie der Bundesregierung. Deutscher Bundestag, Drucksache 19/16075, Berlin

Bundesregierung (2019d): Erster Bericht der Bundesregierung über die Fortschritte bei der Bereitstellung von Daten (1. Open-Data-Fortschrittsbericht). Deutscher Bundestag, Drucksache 19/14140, Berlin

Bundesregierung (2020a): Daten helfen heilen. Innovationsinitiative »Daten für Gesundheit«: Roadmap für eine bessere Patientenversorgung durch Gesundheitsforschung und Digitalisierung. www.bundesgesundheitsministerium.de/fileadmin/Dateien/5_Publikationen/Gesundheit/Berichte/Roadmap_Innovationsinitiative_Daten_fuer_Gesundheit_barrierefrei.pdf (13.12.2021)

Bundesregierung (2020b): Schaffung eines europäischen Cloud- und Datennetzwerkes. Antwort auf die Kleine Anfrage der Fraktion der FDP. Deutscher Bundestag, Drucksache 19/16816, Berlin

Bundesregierung (2021a): Datenstrategie der Bundesregierung. Unterrichtung durch die Bundesregierung. Deutscher Bundestag, Drucksache 19/26450, Berlin

Bundesregierung (2021b): Fünfter Bericht der Bundesregierung über die Fortschritte zur Entwicklung der verschiedenen Felder des Geoinformationswesens im nationalen, europäischen und internationalen Kontext (5. Geo-Fortschrittsbericht). Unterrichtung durch die Bundesregierung. Deutscher Bundestag, Drucksache 19/30737, Berlin

Bundesregierung (2021c): Zukunft des deutschen Traumaregisters. Antwort der Bundesregierung auf die Kleine Anfrage der Abgeordneten der Fraktion BÜNDNIS 90/DIE GRÜNEN. Deutscher Bundestag, Drucksache 19/30638, Berlin

Bundesregierung (2022): Evaluierungsbericht der Bundesregierung gemäß § 142 des Urheberrechtsgesetzes zu den durch das Urheberrechts-Wissensgesellschafts-Gesetz reformierten Vorschriften der §§ 60a bis 60h des Urheberrechtsgesetzes. www.bmj.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/Evaluierungsbericht_Urheberrechts-Wissensgesellschafts-Gesetz.pdf?__blob=publicationFile&v=2 (19.05.2022)

Castelvecchi, D. (2016): Can we open the black box of AI? In: Nature 538, S. 20–23

Cetl, V.; Nunes de Lima, V.; Tomas, R.; Lutz, M.; D'Eugenio, J.; Nagy, A.; Robbrecht, J. (2017): Summary Report on Status of implementation of the INSPIRE Directive in EU. https://publications.jrc.ec.europa.eu/repository/bitstream/JRC109035/jrc109035_jrc109035_jrc_inspire_eu_summaryreport_online.pdf (13.12.2021)

Cheng, J.-Z.; Ni, D.; Chou, Y.-H.; Qin, J.; Tiu, C.-M.; Chang, Y.-C.; Huang, C.-S.; Shen, D.; Chen, C.-M. (2016): Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans. In: Nature Scientific Reports 6, S. 244–254

Ching, T.; Himmelstein, D.; Beaulieu-Jones, B.; Kalinin, A.; Do; B.; Way; G.; Ferrero, E.; Agapow, P.-M.; Zietz, M. et al. (2018): Opportunities and obstacles for deeplearning in biology and medicine. In: Journal of Royal Society Interface 15, Art. 20170387

DEK (Datenethikkommission) (2019): Gutachten der Datenethikkommission der Bundesregierung. www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf?__blob=publicationFile&v=6 (13.12.2021)

Deutscher Ethikrat (2017): Big Data und Gesundheit – Datensouveränität als informationelle Freiheitsgestaltung. www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-big-data-und-gesundheit.pdf (13.12.2021)

DFG (Deutsche Forschungsgemeinschaft) (2013): Sicherung guter wissenschaftlicher Praxis. Denkschrift. www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_1310.pdf (13.12.2021)

DFG (2019): Leitlinien zur Sicherung guter wissenschaftlicher Praxis. www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wissenschaftliche_praxis/kodex_gwp.pdf (13.12.2021)

DGU (Deutsche Gesellschaft für Unfallchirurgie e. V.) (2012): Weißbuch Schwerverletztenversorgung. www.dgu-online.de/fileadmin/dgu-online/Dokumente/6._Versorgung_und_Wissenschaft/Qualit%C3%A4t_und_Sicherheit/20_07_2012_Kurzfassung_Weissbuch_DGU_Auflage2.pdf (13.12.2021)

Dhungel, N.; Carneiro, G.; Bradley, A. (2017): Fully automated classification of mammograms using deep residual neural networks. In: IEEE: 14th International Symposium on Biomedical Imaging, S. 310–314

Dieterich, A.; Braun, B.; Gerlinger, T.; Simon, M. (Hg.) (2019): Geld im Krankenhaus. Eine kritische Bestandsaufnahme des DRG-Systems. Wiesbaden

Dietrich, D. (2011): Was sind offene Daten? Bundeszentrale für politische Bildung, https://www.bpb.de/themen/daten/opendata/64055/was-sind-offene-daten/ (13.12.2021)

Dietzel, J.; Neumann, K.; Glaeske, G.; Greiner, W. (2015): Begleitforschung zum Morbi-RSA (Teil 1). Kriterien, Wirkungen und Alternativen. IGES Institut GmbH, www.iges.com/e6/e1621/e10211/e13470/e13612/e13614/e13616/attr_objs13620/IGES_RSA-Begleitforschung_Teil_1_WEB_ger.pdf (13.12.2021)

DIMDI (Deutsches Institut für Medizinische Dokumentation und Information) (2016): Informationssystem Versorgungsdaten (Datentransparenz). Evaluationsbericht 07/2013–02/2016 Teil 1, Köln

Dirschedl, P. (2012): G-DRG: Ein lernendes System? Vortrag MDK Kongress 2012, https://docplayer.org/22595036-Dr-med-peter-dirschedl-mdk-baden-wuerttemberg-abrechnungspruefungen-des-mdk-im-krankenhaus-buerokratische-last-oder-sinnvolles-korrektiv-g-drg-ein.html (13.12.2021)

Dkfz (Deutsches Krebsforschungszentrum) (2020): Personalisierte Krebstherapie, Präzisionsonkologie, Tumor-Genomsequenzierung. www.krebsinformationsdienst.de/service/iblatt/iblatt-tumor-genomsequenzierung.pdf (13.12.2021)

DKG (Deutsche Krankenhausgesellschaft) (2017): Krankenhäuser als kritische Infrastrukturen – Umsetzungshinweise der Deutschen Krankenhausgesellschaft. www.dkgev.de/fileadmin/default/Mediapool/2_Themen/2.1_Digitalisierung_Daten/2.1.4._IT-Sicherheit_und_technischer_Datenschutz/2.1.4.1._IT-Sicher-heit_im_Krankenhaus/2017_12_19_483_ITSiG_Kritis_Umsetzungshinweise_BSIG_v0.9.pdf (13.12.2021)

DKG (2016): Checkliste zur Erfassung einer familiären Belastung für Brust- und Eierstockkrebs. www.medizin.uni-tuebingen.de/files/view/jReYQ7gOpJW7BgJndB2Droaq/Checkliste%20fam%20Krebserkrankungen.pdf (13.12.2021)

DKG; Deutsche Krebshilfe; AWMF (Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften) (2012): Interdisziplinäre S3-Leitlinie für die Diagnostik, Therapie und Nachsorge des Mammakarzinoms. Version 3.0 Langversion. AWMF-Register Nr. 032 – 0450L, www.leitlinienprogramm-onkologie.de/leitlinien/mammakarzinom/ (13.12.2021)

DKG; Deutsche Krebshilfe; AWMF (2021): Interdisziplinäre S3-Leitlinie für die Früherkennung, Diagnostik, Therapie und Nachsorge des Mammakarzinoms. Langversion 4.4. AWMF-Register Nr. 032 – 0450L, www.leitlinienprogramm-onkologie.de/leitlinien/mammakarzinom/ (13.12.2021)

Döbler, K.; Follert, P. (2021): Stand und Perspektiven der sektorenübergreifen-den Qualitätssicherung. In: Klauber, J.; Wasem, J.; Beivers, A.; Mostert, C. (Hg.): Krankenhausreport 2021. Versorgungsketten – Der Patient im Mittelpunkt. Berlin, S. 223–244

Drechsler, J.; Jentzsch, N. (2018): Synthetische Daten. Innovationspotenziale und gesellschaftliche Herausforderungen. Stiftung Neue Verantwortung, www.stiftung-nv.de/sites/default/files/synthetische_daten.pdf (13.13.2021)

Drepper, D. (2014): Tödliche Keime. https://correctiv.org/recherchen/keime/artikel/2014/11/20/toedliche-keime/ (13.12.2021)

Drösler, S.; Garbe, E.; Hasford, J.; Schubert, I.; Ulrich, V.; van de Ven, W.; Wambach, A.; Wasem, J.; Wille, E. (2017): Sondergutachten zu den Wirkungen des morbiditätsorientierten Risikostrukturausgleich. www.bundesamtsozialesicherung.de/fileadmin/redaktion/Risikostrukturausgleich/20180125Sondergutachten_Wirkung_RSA_2017_korr.pdf (13.12.2021)

EFI (Expertenkommission Forschung und Innovation) (2020): Gutachten zu Forschung, Innovation und technologischer Leistungsfähigkeit Deutschlands 2020. www.e-fi.de/fileadmin/Assets/Gutachten/EFI_Gutachten_2020.pdf (13.12.2021)

EK (Europäische Kommission) (2003): Vorschlag für eine Empfehlung des Rates zur Krebsvorsorge. KOM(2003)230 endgültig, Brüssel

EK (2013): Einführung in das Metadatenmanagement. www.europeandataportal.eu/sites/default/files/d2.1.2_training_module_1.4_introduction_to_metadata_management_de_edp.pdf (13.12.2021)

EK (2020a): Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates über bestreitbare und faire Märkte im digitalen Sektor. COM(2020) 842 final, Brüssel

EK (2020b): Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates über einen Binnenmarkt für digitale Dienste. COM(2020) 825 final, Brüssel

EK (2020c): Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates über europäische Daten-Governance. COM(2020) 767 final Brüssel

EK (2020d): Zur Künstlichen Intelligenz – ein europäisches Konzept für Exzellenz und Vertrauen. Weissbuch. COM(2020) 65 final, Brüssel

EK (2021): Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (Gesetz über künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union. COM(2021) 206 final, Brüssel

EK (2022): Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates über harmonisierte Vorschriften für einen fairen Datenzugang und eine faire Datennutzung. COM(2022) 68 final, Brüssel

Eklund, A.; Nichols, T.; Knutsson, H. (2016): Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. In: PNAS 113(28), S. 7900–7905

Ernst, S. (2010): Google StreetView: Urheber- und persönlichkeitsrechtliche Fragen zum Straßenpanorama. In: Computer und Recht 3, S. 178–184

Ertel, W. (2012): Angewandte Kryptographie. München

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. (1996): From data mining to knowledge discovery in databases. In: AI magazine 17(3), S. 37

FDA (Food and Drug Administration) (2012): Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data – Premarket Notification [510(k)] Submissions. www.fda.gov/media/77635/download (13.12.2021)

FDA (2017): Sentinel Initiative. Final Assessment Report. www.fda.gov/media/107850/download (13.12.2021)

FDP (2019): EU-Medizinprodukteverordnung verantwortungsvoll implementieren – Patientenverordnung sicherstellen. Antrag der Abgeordneten Katrin Helling-Plahr, Christine Aschenberg-Dugnus, Michael Theurer, weiterer Abgeordneter und der Fraktion der FDP. Deutscher Bundestag, Drucksache 19/16035, Berlin

Finlayson, S.; Bowers, J.; Ito, J.; Zittrain, J.; Beam, A.; Ko, I. (2019): Adversarial attacks on medical machine learning. In: Science 363, S. 1287–1289

Forschungsgruppe PMV (2010): Expertise zum Thema: Notwendigkeit des Datenzugangs und der Datentransparenz für ärztliche Körperschaften. Expertise für Bundesärztekammer im Rahmen der Förderinitiative zur Versorgungsforschung. www.bundesaerztekammer.de/fileadmin/user_upload/downloads/Datenzugang-1.pdf (13.12.2021)

Frawley, W.; Piatetsky-Shapiro, G.; Matheus, C. (1992): Knowledge Discovery in Databases: An Overview. In: AI magazine 13(3), S. 57–70

Friedewald, M.; Bieker, F.; Obersteller, H.; Nebel, M.; Martin, N.; Rost, M.; Hansen, M. (2017): Datenschutz-Folgenabschätzung. Ein Werkzeug für einen besseren Datenschutz. www.forum-privatheit.de/download/datenschutz-folgenabschaetzung-3-auflage-2017// (13.12.2021)

Fürstenberg, T.; Laschat, M.; Zich, K.; Klein, S.; Gierling, P.; Nolting, H.; Schmidt, T. (2013): G-DRG-Begleitforschung gemäß § 17b Abs. 8 KHG Endbericht des dritten Forschungszyklus (2008 bis 2010). Berlin

Gamper, J.; Steimann, F. (1996): Medizinische Expertensysteme. Eine kritische Betrachtung. In: APIS Zeitschrift für Politik, Ethik. Wissenschaft und Kultur im Gesundheitswesen, S. 32–40

Gäßner, M. (2002): Expertensysteme (wissensbasierte Systeme) in der Medizin. http://docplayer.org/10915059-Expertensysteme-wissensbasierte-systeme-in-der-medizin-marcus-gaessner.html (13.12.2021)

G-BA (Gemeinsamer Bundesausschuss) (2017): Informationen zum Mammographie-Screening. www.g-ba.de/downloads/17-98-2232/2019-01-21_G-BA_Entscheidungshilfe_Mammographie_bf.pdf? (13.12.2021)

G-BA (2020): Beschluss des Gemeinsamen Bundesausschusses über eine Änderung der Richtlinie Methoden vertragsärztliche Versorgung. Biomarkerbasierte Tests zur Entscheidung für oder gegen eine adjuvante systemische Chemotherapie beim primären Mammakarzinom. www.g-ba.de/downloads/39-261-4524/2020-10-15_MVV-RL_Biomarker-Tests_BAnz.pdf (13.12.2021)

Gehring, P. (2018): Viele Fronten. Forschungsdatenmanagement. In: Forschung & Lehre 9(18), S. 754–756

Geißelmann, K. (2018): Medizinprodukte: Risikoklasse für Apps steigt. In: Deutsches Ärzteblatt 115(12), Art. A538

Gelitz, C. (2019): Künstliche neuronale Netze: Eine neue Form von KI? Spektrum, https://www.spektrum.de/news/eine-neue-form-von-ki/1617834 (13.12.2021)

Gematik (2021): Elektronische Gesundheitskarte und Telematikinfrastruktur. Feature: Verarbeitung von Daten der elektronischen Patientenakte zu Forschungszwecken. https://fachportal.gematik.de/fileadmin/Fachportal/Downloadcenter/Vorabveroeffentlichungen/VorabV_ePA_FDZ/gemF_ePA_FDZ_Anbindung_V1.0.0_CC.pdf (13.12.2021)

Gerlinger, K. (2013): Fernerkundung – Handlungsfelder für einen Technologietransfer in die Länder Afrikas. In: TAB-Brief (42), S. 11–18

Gerste, R. (2014): Das düstere Geheimnis der Pumpe an der Broad Street. Zum 200. Geburtstag von John Snow. In: Chirurgische Allgemeine 15(2), S. 123–126

Gerste, R. (2020): Die Geister des Dr. Snow. In: Zeit online, www.zeit.de/2020/47/cholera-pandemie-john-snow-london-abwasser-infektionskrankheit?utm_referrer=https%3A%2F%2Fwww.google.com%2F (13.12.2021)

Geßner, S.; Dugas, M. (2017): Das Portal für Medizinische Datenmodelle – Der Weg zur strukturierten Dokumentation. In: Forum der Medizin_Dokumentation und Medizin_Informatik 19(1), S. 4–7

Gilbert, F.; Astley, S.; Gillan, M.; Agbaje, O.; Wallis, M.; James, J.; Boggis, C.; Duffy, S. (2008): Single Reading with Computer-Aided Detection for Screening Mammography. In: The New England Journal of Medicine (359), S. 1675–1684

Gilbert, S.; Mehl, A.; Baluch, A.; Cawley, C.; Challiner, J.; Fraser, H.; Millen, E.; Mantazeri, M.; Multmeier, J.; Pick, F.; Richter, C. Türk, E. et al. (2020): How accurate are digital symptom assessment apps for suggesting conditions and urgency advice? A clinical vignettes comparison to GPs. In: British Medical Journal Open

Gillessen-Kaesbach, G.; Paslak, B.; Stehr, F. (2016): Die Seltenen häufiger diagnostizieren. In: Laborwelt 17(4), S. 14

Goldhammer, K.; Wiegand, A. (2017): Ökonomischer Wert von Verbraucherdaten für Adress- und Datenhändler. www.bmjv.de/SharedDocs/Downloads/DE/PDF/Berichte/Oekon_Wert_Daten_Adresshaendler.pdf%3F__blob%3DpublicationFile%26v%3D6 (13.12.2021)

Graham, D.; Campen, D.; Hui, R.; Spence, M.; Cheetham, C.; Levy, G.; Shoor, S.; Ray, W. (2005): Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study. In: Lancet 365(9458), S. 475–481

Grosskreutz, H.; Lemmen, B.; Rüping, S. (2010): Privacy-Preserving Data-Mining. In: Informatik-Spektrum 33(4), S. 380–383

Grüebler, M. (2014): Open Data der Stadt Zürich. Was bringt's? https://de.slideshare.net/Opendatazurich/open-data-der-stadt-zrich-was-bringt-es (13.12.2021)

GSK (GlaxsSmithKline) (2010): Aussetzung der Vermarktung von Arzneimitteln, die Rosiglitazon enthalten (Avandia®, Avandamet® und Avaglim®) in der europäischen Union. www.akdae.de/Arzneimittelsicherheit/RHB/Archiv/2010/20100923.pdf (13.12.2021)

Haas, P. (2017): Elektronische Patientenakten. Einrichtungsübergreifende Elektronische Patientenakten als Basis für integrierte patientenzentrierte Behandlungsmanagement-Plattformen. Bertelsmann Stiftung, www.bertelsmann-stif-tung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/VV_eEPA_Expertise_final.pdf (13.12.2021)

Hallensleben, S.; Hustedt, C.; Fetic; L.; Fleischer, T.; Grünke, P.; Hagendorff, T.; Hauer, M.; Hauschke, A.; Heesen, J.; Herrmann, M.; Hillerbrand, R; Hubig, C. et al. (2020): From Principles to Practice. An interdisciplinary framework to operationalise AI ethics. Bertelsmann Stiftung, www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WKIO_2020_final.pdf (13.12.2021)

He, S. (2020): The Hype of Watson: Why Hasn't AI Taken Over Oncology? www.technologynetworks.com/informatics/articles/the-hype-of-watson-why-hasnt-ai-taken-over-oncology-333571 (13.12.2021)

Heaven, D. (2019): Deep Trouble for Deep Learning. In: Nature 574, S. 163–166

Hecker, D.; Döbel, I.; Petersen, U.; Rauschert, A.; Schmitz, V.; Voss, A. (2017): Zukunftsmarkt Künstliche Intelligenz – Potenziale und Anwendungen. Fraunhofer-Allianz Big Data, www.iais.fraunhofer.de/content/dam/bigdata/de/documents/Publikationen/KI-Potenzialanalyse_2017.pdf (13.12.2021)

Heindel, W.; Bock, K.; Hecht; G.; Heywang-Köbrunner, S.; Kääb-Sanyal, V.; Siegmann-Luz, K.; Weigel, S. (2021): Systematische und qualitätsgesicherte Früher-

kennung des sporadischen Mammakarzinoms. Update Screening-Effekte und wissenschaftliche Studien. In: Der Radiologe 61(2), S. 126–136
Hilty, L.; Oertel, B.; Wölk, M.; Pärli, K. (2012): Lokalisiert und identifiziert. Wie Ortungstechnologien unser Leben verändern. TA-SWISS Band 57, Zürich
Holzinger, A.; Jurisica, I. (Hg.) (2014): Interactive Knowledge Discovery and Data Mining in Biomedical Informatics. State-of-the-Art Survey. Berlin/Heidelberg
Home, P.; Pocock, S.; Beck-Nielsen, H.; Curtis, P.; Gomis, R.; Hanefeld, M.; Jones, N.; Komajda, M.; McMurray, J. (2009): Rosiglitazone evaluated for cardiovascular outcomes in oral agent combination therapy for type 2 diabetes (RECORD): a multicentre, randomised, open-label trial. In: The Lancet 373(9681), S. 2125–2135
Hornung, G. (2018): Industrie 4.0 und das Recht: Drei zentrale Herausforderungen. In: Hornung, G. (Hg.): Rechtsfragen der Industrie 4.0. Datenhoheit – Verantwortlichkeit – rechtliche Grenzen der Vernetzung. Baden-Baden, S. 9–64
Huser, M. (2005): Geo-Informationsrecht. rechtlicher Rahmen für geographische Informationssysteme. Zürich
InEK (Institut für das Entgeltsystem im Krankenhaus) (2020): Weiterentwicklung des G-DRG-Systems für das Jahr 2021. www.g-drg.de/content/download/10160/73513/version/3/file/Abschlussbericht_aG-DRG-System2021.pdf (13.12.2021)
InEK (Institut für das Entgeltsystem im Krankenhaus) (2021): Fallpauschalen-Katalog 2021. www.g-drg.de/aG-DRG-System_2021/Fallpauschalen-Katalog/Fallpauschalen-Katalog_2021 (13.12.2021)
IQWiG (Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen) (2016): Biomarkerbasierte Tests zur Entscheidung für oder gegen eine adjuvante systemische Chemotherapie beim primären Mammakarzinom. IQWiG-Berichte Nr. 457, Köln
IQWiG (2018): Biomarkerbasierte Tests zur Entscheidung für oder gegen eine adjuvante systemische Chemotherapie beim primären Mammakarzinom. Addendum zum Auftrag D14-01. IQWiG-Berichte Nr. 655, Köln
IQWiG (2020): Biomarkerbasierte Tests zur Entscheidung für oder gegen eine adjuvante systemische Chemotherapie beim primären Mammakarzinom. Aktualisierung zum Auftrag D14-01 Rapid Report. IQWiG-Berichte Nr. 883, Köln
Jaume-Palasí, L.; Spielkamp, M. (2017): Ethik und algorithmische Prozesse zu Entscheidungsfindung oder -vorbereitung. AlgorithmWatch, Arbeitspapier Nr. 4, https://algorithmwatch.org/wp-con-tent/uploads/2017/06/AlgorithmWatch_Arbeitspapier_4_Ethik_und_Algorithmen.pdf (13.12.2021)
Jie, Z.; Zhiying, Z.; Li, L. (2021): A meta-analysis of Watson for Oncology in clinical application. In: Nature Scientific Reports 11, Art. 5792
Jones, N. (2014): Computer science: The learning machines. In: Nature 505, S. 146–148
Kääb-Sanyal, V.; Hand, E. (2020): Jahresbericht Evaluation 2018. Deutsches Mammographie-Screening-Programm. Kooperationsgemeinschaft Mammographie, https://fachservice.mammo-programm.de/download/evaluationsberichte/Jahresbericht-Evaluation_2018.pdf (13.12.2021)
Kaplan, S. (2017): Failure to warn: An early warning system for drug risks falls flat. www.statnews.com/2017/06/06/sentinel-fda-drug-risks/ (13.12.2021)
Karg, M. (2008): Datenschutzrechtliche Rahmenbedingungen für die Bereitstellung von Geodaten für die Wirtschaft. Unabhängiges Landeszentrum für Datenschutz

Schleswig-Holstein, www.datenschutzzentrum.de/uploads/geodaten/datenschutzrechtliche-rahmenbedingungen-bereitstellung-geodaten.pdf (13.12.2021)
Karlberg, S. (2018): Gene und Daten gegen Krebs. Tagesspiegel, https://www.tagesspiegel.de/wissen/gene-und-daten-gegen-krebs-5517127.html
KBV (Kassenärztliche Bundesvereinigung) (2016): Qualitätssicherungsvereinbarung MRSA. www.kbv.de/media/sp/QS-MRSA.pdf (13.12.2021)
KBV (2021): Richtlinie der Kassenärztlichen Bundesvereinigung nach § 75 Abs. 7 SGB V zur Vergabe der Arzt-, Betriebsstätten- sowie der Praxisnetznummern. www.kbv.de/media/sp/Arztnummern_Richtlinie.pdf (13.12.2021)
Kettritz, U. (2018): Überdiagnose im Mammographie-Screening. aktuelle Daten und Bewertung. www.ggg-b.de/vortraege/15.pdf (13.12.2021)
Enquete-Kommission (Enquete-Kommission Künstliche Intelligenz) (2020): Bericht der Enquete-Kommission Künstliche Intelligenz – Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale. Deutscher Bundestag, Drucksache 19/23700, Berlin
Knobloch, B.; Weidner, J. (2000): Eine kritische Betrachtung von Data-Mining-Prozessen. Ablauf, Effizienz und Unterstützungspotenziale. In: Jung, R.; Winter, R. (Hg.): Data Warehousing 2000. Methoden, Anwendungen, Strategien. Heidelberg, S. 345–365
Köhler, S.; Schulz, M.; Krawitz, P.; Bauer, S.; Dölken, S.; Ott, C.; Mundlos, C.; Horn, D.; Mundlos, S.; Robinson, P. (2009): Clinical diagnostics in human genetics with semantic similarity searches in ontologies. In: American Journal of Human Genetics 85(4), S. 457–464
Köhler, S.; Vasilevsky, N.; Engelstad, M.; Foster, E.; McMurray, J.; Aymé, S.; Bayman, G.; Bello, S.; Boerkoel, C.; Boycott, M.; Brudno, M.; Buske, O. (2017): The Human Phenotype Ontology in 2017. In: Nucleic Acids Research 45, S. 865–876
Kooi, T.; Litjens, G.; van Ginneken, B.; Gubern-Mérida, A.; Sánchez, C.-I.; Mann, R.; den Heeten, A.; Karssemeijer, N. (2017): Large scale deep learning for computer aided detection of mammographic lesions. In: Medical Image Analysis 35, S. 303–312
Krahnert, S. (2016): Aktuelle Debatte: Aufweichung der ärztlichen Schweigepflicht? www.krahnert-medizinrecht.de/debatte-aufweichung-schweigepflicht/ (13.12.2021)
Kreutzer, T.; Fischer, G. (2022): Das Urheberrechts-Wissensgesellschafts-Gesetz in der Praxis. https://irights.info/wp-content/uploads/2022/05/Kurzstudie-Das-UrhWissG-in-der-Praxis_Publikation_DOI.pdf (19.05.2022)
Kriegel, H.-P.; Kröger, P.; Zimek, A. (2009): Outlier Detection Techniques. www.dbs.ifi.lmu.de/Publikationen/Papers/tutorial_slides.pdf (13.12.2021)
Krokotsch, A. (2017): Erst Papierberg, dann Datenmüll: Fallstricke der digitalen Patientenakte. In: Forum der Medizin_Dokumentation und Medizin_Informatik 19(1), S. 14–17
Ksoll, W.; Schiedbauer, T.; Beck, A. (2017): Open Data – Wertschöpfung im digitalen Zeitalter. Bertelsmann Stiftung, www.bertelsmann-stiftung.de/fileadmin/files/Projekte/Smart_Country/OpenData_2017_final.pdf (13.12.2021)
Kummer, K.; Pischler, N.; Zeddies, W. (2006): Das Amtliche deutsche Vermessungswesen. Stark in den Regionen und einheitlich im Bund – für Europa. In: Zeitschrift für Geodäsie, Geoinformation und Landmanagement 131(5), S. 234–241

Kunze, A. (2013): Patientendaten. Behandeln statt verkaufen. Zeit online, https://www.zeit.de/2013/48/datenschutz-patientendaten?utm_referrer=https%3A%2F%2Fwww.google.com%2F (13.12.2021)
Kuschel, L. (2018): Wem »gehören« Forschungsdaten? Zur Rechtslage nach Urheber- und Datenschutzrecht. In: Forschung & Lehre 9(18), S. 764–766
Kuzev, P. (2016): Open Data: Die wichtigsten Fakten zu offenen Daten. Konrad-Adenauer-Stiftung, www.kas.de/wf/doc/kas_44530-544-1-30.pdf?160315122244 (13.12.2021)
Laursen, L. (2016): Doctors Still Struggle to Make the Most of Computer-Aided Diagnosis. https://spectrum.ieee.org/doctors-still-struggle-to-make-the-most-of-computer-aided-diagnosis (13.12.2021)
Lehman, C.; Wellman, R.; Buist, D.; Kerlikowske, K.; Tosteson, A.; Miglioretti, D.; Breast Cancer Surveillance Consortium (2015): Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. In: JAMA Intern Med. 175(11), S. 1828–1837
Lenzen-Schulte, M. (2017): Medizinische Suchmaschinen: Mit einem Mausklick zur Diagnose. In: Deutsches Ärzteblatt 114(25), S. A1231–A1233
Liu, X.; Faes, L.; Kale, A.; Wagner, S.; Fu, D.; Bruynseels, A.; Mahendiran, T.; Moraes, G.; Shamdas, M.; Kern, C.; Ledsam. J.; Schmid, M. et al. (2019): A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. In: The Lancet Digital Health 1(6), S. E271–E291
Lotter, W.; Sorensen, G.; Cox, D. (2017): A Multi-Scale CNN and Curriculum Learning Strategy for Mammogram Classification. In: Cardoso, J.; Arbel, T.; Carneiro, G.; Syeda-Mahmood, T.; Moradi, M.; Bradley, A.; Greenspan, H.; Papa, J.; Madabushi, A.; Nascimento, J.; Cardoso, J. et al. (Hg.): Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Cham, S. 169–177
Lücker, V. (2018): Medizinproduktrechtliche Rahmenbedingungen für E-Health-Produkte im europäischen Wirtschaftsraum. In: Bundesgesundheitsblatt 61(1), S. 278–284
Lüdtke, H. (2002): Tübinger Mole Analyzer: Digitale Bildanalyse für die Melanomfrüherkennung. In: Deutsches Ärzteblatt 99(41), S. 17–18
Machotta, T. (2013): Verkauf von Rezeptdaten? Apotheken-Rechenzentrum weist Vorwürfe zurück. apotheken.de, www.apotheken.de/index.php?id=25&cv=nc%3F&tx_ttnews%5Btt_news%5D=9380&cHash=48bcd2c8389acb16df217926e442c62f (13.12.2021)
Matthews, R. (2000): Storks Deliver Babies (p=0.008). In: Teaching Statistics 22(2), S. 36–38
McKee, L. (2012): OGC History. www.opengeospatial.org/ogc/historylong (13.12.2021)
McKinney, S.; Sieniek, M.; Godbole, V.; Godwin, J.; Antropova, N.; Ashrafian, H.; Back, T.; Chesus, M.; Corrado, G.; Darzi, A.; Etemadi, M.; Garcia-Vicente, F. et al. (2020): International evaluation of an AI system for breast cancer screening. In: Nature 577, S. 89–113
Mendes, R.; Vilela, J. (2017): Privacy-Preserving Data Mining: Methods, Metrics, and Applications. In: IEEE Access 5, S. 10562–10582
Metschke, R.; Wellbrock, R. (2002): Datenschutz in Wissenschaft und Forschung. Berliner Beauftragter für Datenschutz und Informationsfreiheit, www.forschungsdaten-bildung.de/files/metschkewellbrock2002.pdf (13.12.2021)

MII (Medizininformatik Initiative) (2020): Mustertext Patienteneinwilligung. Version 1.6.d. www.medizininformatik-initiative.de/sites/default/files/2020-04/MII_AG-Consent_Einheitlicher-Mustertext_v1.6d.pdf (13.12.2021)
Montjoye, Y.-A. de; Radaelli, L.; Singh, V.; Pentland, A. (2015): Unique in the shopping mall: On the reidentifiability of credit card metadata. In: Science 347(6221), S. 536–539
Moreau, C.; Wolfsteiner, M. (2017): Anonymisierungsverfahren in der Kommunalstatistik: Theorie und praktische Anwendung. In: Frankfurter Statistische Berichte, S. 48–55
Mühr, C. (2018): Was fehlt für eine sektoren- und einrichtungsübergreifende Qualitätssicherung? www.g-ba.de/downloads/17-98-4690/2018-09-24_QS-Konferenz_PV1-2_Mu%CC%88hr_G-BA.pdf (13.12.2021)
Müller-Quade, J.; et al. (2020): Sichere KI-Systeme für die Medizin. Whitepaper aus der Plattform Lernende Systeme. www.plattform-lernende-syste-me.de/files/Downloads/Publikationen/AG3_6_Whitepaper_07042020.pdf (13.12.2021)
Muscholl, M.; Kadioglu, D.; Lablans, M.; Storf, H.; Göbel, J.; Pfalz, A.; Ückert, F.; Wagner, T. (2016): OSSE – Open-Source-Registersystem für Seltene Erkrankungen in der EU. www.osse-register.de/OSSE_summary_de.pdf (13.12.2021)
Nielsen, M. (2018): Künstliche Intelligenz. Alpha Go – Computer lernen Intuition. In: Spektrum der Wissenschaft 1, S. 22–27
Nissen, S.; Wolski, K. (2007): Effect of Rosiglitazone on the Risk of Myocardial Infarction and Death from Cardiovascular Causes. In: The New England Journal of Medicine 356, S. 2457–2471
Nohr, H. (2017): Big Data im Lichte der EU-Datenschutz-Grundverordnung. In: JurPC Web-Doc. 111, Abs. 1–86
Orwat, C. (2019): Diskriminierungsrisiken durch Verwendung von Algorithmen. Antidiskriminierungsstelle des Bundes, https://publikationen.bibliothek.kit.edu/1000103134 (13.12.2021)
Paik, S.; Shak, S.; Tang, G.; Kim, C.; Baker, J.; Cronin, M.; Baehner, F.; Walker, M.; Watson, D.; Park, T.; Hiller, W.; Fisher, E. et al. (2004): A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer. In: New England Journal of Medicine 351(27), S. 2817–2826
Pasquinelli, M. (2018): Matadata Society. In: Braidotti, R.; Hlavajova, M. (Hg.): Posthuman Glossary. London, S. 253–256
Pfeffer, S. (2002): Expertensysteme in der Medizin. In: JurPC Web-Dok. 91, Abs. 1–23
Phillips, F. (2017): A perspective on »Big Data«. In: Science and Public Policy 44(5), S. 730–737
Pisano, E. (2020): AI shows promise for breast cancer screening. In: Nature 577, S. 35–36
Prognos (2016): Gutachten zum aktuellen Umsetzungsstand des KFRG. Basel, www.prognos.com/sites/default/files/2021-01/20160818_prognos_schlussversion-gutachten_kkr.pdf (13.12.2021)
Prognos (2018): Stand der klinischen Krebsregistrierung. Basel, www.gkv-spitzenverband.de/media/dokumente/krankenversicherung_1/qualitaetssicherung_2/klinische_krebsregister/2018-10-19_Prognos-Gutachten_Stand_der_klinischen_Krebsregistrierung_final.pdf (13.12.2021)

Puppe, F. (2014): Entscheidungsunterstützungssyteme, medizinische. Gesellschaft für Informatik, https://gi.de/informatiklexikon/entscheidungsunterstuetzungssytememedizinische (13.12.2021)
Rahm, E. (2015): Vorlesungsskript. 1. Data Warehouses – Einführung. https://dbs.unileipzig.de/file/dw-kap1.pdf (13.12.2021)
RatSWD (Rat für Sozial- und WirtschaftsDaten) (2017): Forschungsethische Grundsätze und Prüfverfahren in den Sozial- und Wirtschaftswissenschaften. www.ratswd.de/dl/RatSWD_Output9_Forschungsethik.pdf (13.12.2021)
RatSWD (2020): Datenerhebung mit neuer Informationstechnologie. Empfehlungen zu Datenqualität und -management, Forschungsethik und Datenschutz. www.ratswd.de/dl/RatSWD_Output6.6_Datenerhebung-neueIT.pdf (13.12.2021)
Revermann, C.; Sauter, A. (2007): Biobanken für die humanmedizinische Forschung und Anwendung. Studien des Büros für Technikfolgen-Abschätzung beim Deutschen Bundestag 23, Berlin
Rey, G.; Wender, K. (2018): Neuronale Netze. Eine Einführung in die Grundlagen, Anwendungen und Datenauswertung. Bern
RfII (Rat für Informationsinfrastrukturen) (2020): Stellungnahme des RfII. Datentreuhandstellen gestalten – Zu Erfahrungen der Wissenschaft. www.rfii.de/?wpdmdl=4259 (13.12.2021)
Rhön-Klinikum (2017): Kognitives Assistenzsystem für Einsatz im Krankenhaus erfolgreich erprobt – Rhön-Klinikum AG. www.rhoen-klinikum-ag.com/presse/pressemeldungen/news/article/rhoen-klinikum-ag-kognitives-assistenzsystemfuer-einsatz-im-krankenhaus-erfolgreich-erprobt.html (13.12.2021)
Rhön-Klinikum; IBM (2016): IBM-Watson-Technologie: Uniklinik Marburg testet kognitiven Assistenten für die Diagnose seltener Krankheiten. E-HEALTH-COM, https://e-health-com.de/details-news/ibm-watson-technologie-uniklinik-marburgtestet-kognitiven-assistenten-fuer-die-diagnose-seltener-kr/ (13.12.2021)
Ribli, D.; Horváth, A.; Unger, Z.; Pollner, P.; Csabai, I. (2018): Detecting and classifying lesions in mammograms with Deep Learning. In: Nature Scientific Reports 8, Art. 4165
Riehm, T. (2018): Dateneigentum – Schutz nach allgemeinem Zivilrecht. In: Hornung, G. (Hg.): Rechtsfragen der Industrie 4.0. Datenhoheit – Verantwortlichkeit – rechtliche Grenzen der Vernetzung. Baden-Baden, S. 73–96
RKI (Robert Koch Institut) (2016): Regionale Verteilung des Anteils von MRSA und VRE bei nosokomialen Infektionen mit S. aureus und Enterokokken Untersuchung auf Intensivstationen sowie bei postoperativen Wundinfektionen. In: Epidemiologisches Bulletin 22, S. 191–196
Ross, C.; Swetlitz, I. (2017): IBM pitched Watson as a revolution in cancer care. It's nowhere close. STAT, www.statnews.com/2017/09/05/watson-ibm-cancer/ (13.12.2021)
Rott, P. (2018): Rechtspolitischer Handlungsbedarf im Haftungsrecht, insbesondere für digitale Anwendungen. Gutachten im Auftrag des Verbraucherzentrale Bundesverbandes. Verbraucherzentrale Bundesverband, www.vzbv.de/sites/default/files/downloads/2018/05/04/gutachten_handlungsbedarf_im_haftungsrecht.pdf (13.12.2021)
Rüchardt, D. (2019): Chancen und Risiken der Plattformökonomie. Wie Plattformen die Digitalwirtschaft bestimmen. Computerwoche, www.cowo.de/a/3547305 (13.12.2021)

Rüschemeyer, G. (2020): Die mangelnde Offenlegung von Studienergebnissen kann Ihrer Gesundheit schaden! Cochrane Deutschland, www.cochrane.de/de/news/die-mangelnde-offenlegung-von-studienergebnissen-kann-ihrer-gesundheit-schaden (13.12.2021)

Samek, W.; Montavon, G.; Vedaldi, A.; Hansen, L.; Müller, K. (Hg.) (2019): Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Cham

Schink, T.; Holstiege, J.; Kowalzik, F.; Zepp, F.; Garbe, E. (2014): Risk of febrile convulsions after MMRV vaccination in comparison to MMR or MMR+V vaccination. In: Vaccine 32, S. 645–650

Schneider, U. (2015): Sekundärnutzung klinischer Daten – rechtliche Rahmen-bedingungen. Schriftenreihe der TMF 12, Berlin

Scholz, R.; Beckedahl, M.; Noller, S.; Renn, O. (Hg.) (2021): DiDaT Weißbuch. Verantwortungsvoller Umgang mit digitalen Daten – Orientierungen eines transdisziplinären Prozesses. Baden-Baden

Schröder, G. (2003): Mut zum Frieden und Mut zur Veränderung. Regierungserklärung von Bundeskanzler Gerhard Schröder vor dem Deutschen Bundestag am 14. März 2003 in Berlin. www.bundestag.de/webarchiv/textarchiv/2013/43257637 (13.12.2021)

Schubert, I.; Ihle, P.; Köster, I.; Küpper-Nybelen, J.; Rentzsch, M.; Stallmann, C.; Swart, E.; Winkler, C. (2014): Daten für die Versorgungsforschung. Zugang und Nutzungsmöglichkeiten. https://e-health-com.de/fileadmin/user_upload/dateien/Downloads/dimdi-sekundaerdaten-expertise.pdf (13.12.2021)

Schüller-Zwierlein, A.; Leiwesmayer, B. (2018): Neuerungen im Urheberrecht. Stand und Perspektiven. www.uni-regensburg.de/bibliothek/medien/pdf/urheberrecht_neuerungen.pdf (13.12.2021)

Shearer, C. (2000): The CRISP-DM Model: The New Blueprint for Data Mining. In: Journal of Data Warehousing 5, S. 13–20

Shortliffe, E. (1987): Computer programs to support clinical decision making. In: Journal of the American Medical Association 258(1), S. 61–66

Sickles, E.; D'Orsi, C.; Bassett, L.; Appleton, C.; Berg, W.; Burnside, E.; Feig, S.; Gavenonis, S.; Newell, M.; Trinh, M. et al. (2013): ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System. 5th Edition. American College of Radiology, Reston

Siemoneit, O. (2018): Big Data, quo vadis? Trends, Treiber, Determinanten, Wildcards. Karlsruher Institut für Technologie, Scientific Working Papers 86, https://publikationen.bibliothek.kit.edu/1000082069 (13.12.2021)

Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; Lillicrap, T. et al. (2017): Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. https://arxiv.org/pdf/1712.01815 (13.12.2021)

SPD; BÜNDNIS 90/DIE GRÜNEN (2001): Nutzung von Geoinformationen in der Bundesrepublik Deutschland. Entschließungsantrag der Abgeordneten Dr. Margrit Wetzel, Klaus Barthel (Starnberg), Dr. Axel Berg, weiterer Abgeordneter und der Fraktion der SPD sowie der Abgeordneten Hans-Josef Fell, Kerstin Müller (Köln), Rezzo Schlauch und der Fraktion BÜNDNIS 90/DIE GRÜNEN zu der Großen Anfrage der Abgeordneten Dr.-Ing. Rainer Jork, Ilse Aigner, Günter Baumann, weiterer Abgeordneter und der Fraktion der CDU/CSU – Drucksachen 14/3214, 14/4139 –. Deutscher Bundestag, Drucksache 14/5323, Berlin

Stausberg, J.; Semler, S.; Neugebauer, E. (2014): Ein Register von Registern und Kohorten: Das Registerportal von TMF und DNVF. In: GMDS: 59. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V., Göttingen

Stock, W. (2018): Informationsmarkt. Heinrich-Heine-Universität Düsseldorf, https://www.ling.hhu.de/fileadmin/redaktion/Fakultaeten/Philosophische_Fakultaet/Sprache_und_Information/Informationswissenschaft/Dateien/Wolfgang_G._Stock/Informationsmarkt.pdf (13.12.2021)

SVR Gesundheit (Sachverständigenrat zur Begutachtung der Entwicklung im Gesundheitswesen) (2014): Bedarfsgerechte Versorgung − Perspektiven für ländliche Regionen und ausgewählte Leistungsbereiche. www.svr-gesundheit.de/fileadmin/Gutachten/Gutachten_2014/Langfassung2014.pdf (13.12.2021)

SVR Gesundheit (2021): Digitalisierung für Gesundheit – Ziele und Rahmenbedingungen eines dynamisch lernenden Gesundheitssystems. www.svr-gesundheit.de/gutachten/gutachten-2021/ (13.12.2021)

TAB (Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag) (2010): Stand und Perspektiven klinischer Forschung in Deutschland und im Vergleich zu anderen Ländern und besonderer Berücksichtigung nichtkommerzieller Studien. (Bührlen, B.; Georgieff, P.; Vollmar, H.) TAB-Arbeitsbericht Nr. 135, Berlin

TAB (2012): Fernerkundung: Anwendungs-potenziale in Afrika. (Gerlinger, K.) TAB-Arbeitsbericht Nr. 154, Berlin

TAB (2014): Big Data in der Cloud. (Leimbach, T.; Bachlechner, D.) TAB-Hintergrundpapier Nr. 18, Berlin

TAB (2016): Technologien und Visionen der Mensch-Maschine-Entgrenzung. (Kehl, C.; Coenen, C.) TAB-Arbeitsbericht Nr. 167, Berlin

TAB (2017a): Neue Arzneimittel gegen vernachlässigte Krankheiten. (Gerlinger, K.) TAB-Arbeitsbericht Nr. 170, Berlin

TAB (2017b): Microtargeting: psychometrische Analyse mittels Big Data. (Kind, S.; Weide, S.) TAB-Themenkurzprofil Nr. 18, Berlin

TAB (2018): Gesundheits-Apps. (Evers-Wölk, M.; Oertel, B.; Sonk, M.) TAB-Arbeitsbericht Nr. 179, Berlin

TAB (2020): Mögliche Diskriminierung durch algorithmische Entscheidungssysteme und maschinelles Lernen – ein Überblick. (Kolleck, A.; Orwat, C.) TAB-Hintergrundpapier Nr. 24, Berlin

TAB (2021): Digitalisierung der Landwirtschaft. (Kehl, C.; Steiger, S.) TAB-Arbeitsberichte Nr. 193 und 194, Berlin

TAB (2022a): Beobachtungstechnologien im Bereich der zivilen Sicherheit – Möglichkeiten und Herausforderungen. (Caviezel, C.; Hempel, L.; Revermann, C.; Steiger, S.) TAB-Arbeitsbericht Nr. 190, Berlin

TAB (2022b): Innovative Technologien, Prozesse und Produkte in der Bauwirtschaft. (Kehl, C.; Achternbosch, M.; Revermann, C.) TAB-Arbeitsbericht Nr. 199, Berlin

TAB (2022c): Künstliche Intelligenz und Distributed-Ledger-Technologie in der öffentlichen Verwaltung. Ein Überblick von Chancen und Risiken einschließlich der Darstellung international einschlägiger Praxisbeispiele. (Evers-Wölk, M.; Kluge, J.; Steiger, S.) TAB-Arbeitsbericht Nr. 201, Berlin

Taichman, D.; Backus, J.; Baethge, C.; Bauchner, H.; de Leeuw, P.; Drazen, J.; Fletcher, J.; Frizelle, F.; Groves, T.; Haileamlak, A.; James, A.; Laine, C. et al. (2016):

Bereitstellung von Primärdaten klinischer Studien. Ein Vorschlag des International Committee of Medical Journal Editors. In: Deutsches Ärzteblatt 113(4), S. 41–43

Taylor, P.; Potts, H. (2008): Computer aids and human second reading as interventions in screening mammography: Two systematic reviews to compare effects on cancer detection and recall rate. In: European Journal of Cancer 44(6), S. 798–807

Tenckhoff, B. (2015): Nutzung von Daten für die einrichtungs- und sektorenübergreifende Qualitätssicherung. Vortrag 60. GMDS-Jahrestagung, Krefeld

Tenckhoff, B. (2017): Sekundärdatennutzung medizinischer Dokumentation. Chancen und Risiken ambulanter Routinedaten im KV-System. In: Forum der Medizin_Dokumentation und Medizin_Informatik 1, S. 11–14

Thiel, R.; Deimel, L.; Schmidtmann, D.; Piesche, K.; Hüsing, T.; Rennoch, J.; Stroetmann, V.; Stroetmann, K. (2018): #SmartHealthSystems. Digitalisierungsstrategien im internationalen Vergleich. Bertelsmann Stiftung, www.bertelsmann-stiftung.de/fileadmin/files/Projekte/Der_digitale_Patient/VV_SHS-Gesamtstudie_dt.pdf (13.12.2021)

Timmers, J.; den Heeten, G.; Adang, E.; Otten, J.; Verbeek, A.; Broeders, M. (2012): Dutch digital breast cancer screening: implications for breast cancer care. In: European journal of public health 22(6), S. 925–929

Tremmel, S.; Gieselmann, H.; Eikenberg, R. (2019): Datenschleuder. Massive Datenschutzmängel in der Gesundheits-App Ada. In: c't magazin für computertechnik 22, S. 14–15

Triaille, J.-P.; de Meeus d'Argenteuil, J.; de Francquen, A. (2014): Study on the legal framework of text and data mining (TDM). De Wolf & Partners, https://data.europa.eu/doi/10.2780/1475 (13.12.2021)

von Unger, H.; Simon, D. (2016): Ethikkommissionen in den Sozialwissenschaften. Historische Entwicklungen und internationale Kontroversen. RatSWD Working Papers Nr. 253, www.ratswd.de/dl/RatSWD_WP_253.pdf (13.12.2021)

van 't Veer, L.; Hongyue, D.; van de Vijver, M.; He, Y.; Hart, A.; Mao, M.; Peterse, H.; van der Kooy, K.; Marton, M.; Witteveen, A.; Schreiber, G. et al. (2002): Gene expression profiling predicts clinical outcome of breast cancer. In: Nature 415, S. 530–536

Veta, M.; Pluim, J.; van Diest, P.; van de Viergever, M. (2014): Breast Cancer Histopathology Image Analysis: A Review. In: IEEE Transactions on Biomedical Engineering 61(5), S. 1400–1411

Vfa (Verband der forschenden Arzneimittelhersteller) (2020): Stellungnahme zum Referentenentwurf des Bundesministeriums für Gesundheit. Verordnung zur Neufassung der Datentransparenzverordnung und zur Änderung der Datentransparenz-Gebührenverordnung. www.vfa.de/download/stellungnahme-referentenentwurf-datentransparenzverordnung.pdf (13.12.2021)

Vodafone Institute for Society and Communications (Hg.) (2016): Big Data: Wann Menschen bereit sind, ihre Daten zu teilen. www.vodafone-institut.de/wp-content/uploads/2016/01/VodafoneInstitute-Survey-BigData-Highlights-de.pdf (13.12.2021)

Vomweg, T. (2017): Lektionen vom Mammographie-Screening. In: Pneumologe 14(3), S. 131–139

Von der Weiden, S. (2018): KI stellt die bessere Diagnose als der Arzt. In: VDI nachrichten 18, S. 13

Wangermann, T. (2016): Open Data aus internationaler Perspektive. www.data.gv.at/wp-content/uploads/2016/07/Open-Data-aus-internationaler-Perspektive.pdf (13.12.2021)
Ware, W. (2005): The Vioxx Saga: Perspective on the Recall. www.yourhealthbase.com/Vioxx.htm (13.12.2021)
Weichert, T. (2018): Big Data im Gesundheitsbereich. www.abida.de/sites/default/files/ABIDA%20Gutachten-Gesundheitsbereich.pdf (13.12.2021)
Welzel, C.; Grosch, D. (2018): Das ÖFIT-Trendsonar Künstliche Intelligenz. Kompetenzzentrum Öffentliche IT, www.oeffentliche-it.de/documents/10181/14412/Das+ÖFIT-Trendsonar+Künstliche+Intelligenz (13.12.2021)
Wiegerling, K.; Nerurkar, M.; Wadephul C. (2018): Ethische und anthropologische Aspekte der Anwendung von Big-Data-Technologien. In: Kolany-Raiser, B.; Heil, R.; Orwat, C.; Hoeren, T. (Hg.): Big Data und Gesellschaft. Eine multidisziplinäre Annäherung. Wiesbaden, S. 1–68
Wilkens, L. (2017): Methodenstreit auf dem Rücken von Brustkrebspatientinnen? Die Versorgungsrealität. Klinikum Region Hannover, www.hello-healthcare.com/files/site-files/Events/01%20Summit%2010.%20Maerz%202017/02%20Vortraege/5.%20Prof.%20Wilkens.pdf (13.12.2021)
Wilkinson, M.; Dumontier, M.; Aalbersberg, I.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; Bonino da Silva Santos, L.; Bourne, P.; Bouwman, J. et al. (2016): The FAIR Guiding Principles for scientific data management and stewardship. In: Nature Scientific Data 3, Art. 160018
Winsberg, F.; Elkin, M.; Marcey, J.; Bordaz, V.; Weymouth, W. (1967): Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis. In: Radiology 89(2), S. 211–215
WD (Wissenschaftliche Dienste) (2016): Zur Diskussion eines Patientenentschädigungs- und Härtefallfonds. Deutscher Bundestag, Dokumentation Nr. WD 9-3000-043/16, Berlin
WD (2019): E-Government in Deutschland. Aktueller Stand auf Bundes- und Landesebene. Deutscher Bundestag, Sachstand Nr. WD 3-3000-134/19, Berlin
Witten, I.; Frank, E.; Hall, M. (2011): Data Mining: practical machine learning tools and techniques. Burlington
WR (Wissenschaftsrat) (2009): Stellungnahme zum Status und der zukünftigen Entwicklung des Rates für Sozial- und Wirtschaftsdaten (RatSWD). Drs. 9504-09, Aachen
Zweig, K. (2016): 1. Arbeitspapier: Was ist ein Algorithmus? Algorithm Watch, http://algorithmwatch.org/arbeitspapier-was-ist-ein-algorithmus/ (13.12.2021)
Zweig, K. (2019a): Algorithmische Entscheidungen: Transparenz und Kontrolle. In: Arnold, N.; Wangermann, T. (Hg.): Digitalisierung und künstliche Intelligenz: Orientierungspunkte. Konrad-Adenauer-Stiftung e. V., Berlin, S. 143–163
Zweig, K. (2019b): Algorithmische Entscheidungen: Transparenz und Kontrolle. Konrad-Adenauer-Stiftung, Analysen & Argumente Nr. 338, Berlin

8 Anhang

8.1 Data-Mining im Urheber- und Leistungsschutzrecht

Die Begriffe *Text- und Data-Mining* (TDM) sind seit 2018 im deutschen und seit 2019 im europäischen Urheberrecht verankert. Das Urheberrecht definiert geistige bzw. immaterielle Eigentumsrechte, also die Rechte urhebender Personen an ihren Texten, Bildern, Datensammlungen (Werken), deren Reichweite und Grenzen sowie deren Verwertungsmöglichkeiten. Text- und Data-Mining wird dort definiert als automatisierte Analyse von digitalen Werken bzw. von Texten und Daten in digitaler Form um daraus Informationen insbesondere über Muster, Trends und Korrelationen zu gewinnen (§ 44b UrhG[168]; Richtlinie 2019/790/EU). Diese Definition entspricht der Auffassung von Data-Mining im engeren Sinn innerhalb dieses Berichts.

Schutzgegenstand und immaterielle Eigentumsrechte

Urheberrechtlich geschützt sind *geistig schöpferische Werke*, darunter fallen Darstellungen wissenschaftlicher oder technischer Art wie Karten oder Tabellen, (Licht-)Bilder und Filme, Sprach- und Schriftwerke sowie Computerprogramme (§ 2 UrhG), Datenbanken (§ 4 UrhG) oder nichtamtliche Normen (z. B. ISO- oder DIN-Normen), nicht aber amtliche Werke wie Gesetze, amtliche Erlasse, Bekanntmachungen sowie Entscheidungen und Leitsätze dazu (§ 5 UrhG). Voraussetzung für einen Urheberrechtsschutz ist ein Mindestmaß an schöpferischer Tätigkeit bzw. geistig kreativem Schaffen, was über technisch-handwerkliche Tätigkeiten und Fleißarbeit, wie die Bedienung technischer Geräte zur Datenerhebung, Dokumentationen, systematisches Aufzählen oder Klassifizierungstätigkeiten anhand sachlogischer Kriterien, hinausgeht (Huser 2005, S. 72 ff.; Kuschel 2018; Schepers et al. 2015, S. 259). Geschützt wird die Form eines Werkes, die zugrunde liegende spezifische kreative Idee und deren Ausgestaltung, nicht jeder einzelne Inhaltsbestandteil: Bei wissenschaftlichen Darstellungen, Karten oder Plänen sind nicht die zugrundeliegenden einzelnen Werte/Daten (die in der Regel maschinell erzeugt und nicht kreativ geschaffen wurden) oder Worte (Zitationen kleinerer Bestandteile sind ohne Erlaubnis der urhebenden Person zulässig, sofern diese genannt wird), bei Bildern ist nicht jeder einzelne Bildpunkt, bei Tabellen nicht jede einzelne Ziffer geschützt. Auch bei Datenbanken werden nur die kreative Auswahl und die Anordnung von Daten, nicht aber einzelne Inhaltselemente urheberrechtlich geschützt.

168 Gesetz über Urheberrecht und verwandte Schutzrechte (Urheberrechtsgesetz – UrhG)

Das Urheberrecht gewährt natürlichen Personen geistige Eigentumsrechte (Intellectual Property Rigths) an ihrem Werk. Sie umfassen

- *unveräußerliche Persönlichkeitsrechte* (automatische Urheberschaftsanerkennung bei der Veröffentlichung eines Werkes; §§ 12 ff. UrhG) und
- *veräußerliche Verwertungsrechte* (u. a. zur Vervielfältigung, Veröffentlichung, Bearbeitung, Umgestaltung, freien Benutzung; §§ 15 ff. UrhG), die 70 Jahre nach dem Tod der urhebenden Person(en) erlöschen; danach ist das jeweilige Werk gemeinfrei für alle verwendbar (§ 64 UrhG).

Strukturierte, semantisch und syntaktisch normierte (Roh-)Datensätze und -bestände, die das wesentliche Fundament für Data-Mining-Prozesse im Sinne dieses Berichts sind, entstehen in der Regel nicht durch schöpferische, sondern durch technische Tätigkeiten und Fleißarbeit, werden zunehmend automatisiert erzeugt und in standardisierten Datenbanken gehalten. Diese datengenerierenden und -aufbewahrenden Leistungen werden wegen der zu geringen Schöpfungshöhe allerdings nicht urheberrechtlich, sondern von verwandten Schutzrechten, den Leistungsschutzrechten, erfasst (Teil 2 UrhG). Derartige Datensätze und -bestände werden nur mittels veräußerlicher Verwertungsrechte geschützt. Für unterschiedliche Leistungsformen definiert das Urhebergesetz jeweils spezifische Schutzfristen, die im Lauf der Zeit ausgeweitet wurden. Die Spanne reicht von 15 Jahren für Datenbanken als systematische Anordnung von Werken, Daten oder anderen unabhängigen Elementen (§§ 87a ff. UrhG) über 50 Jahre für Licht- und Laufbilder sowie Filme (§§ 72 und 94 f. UrhG) bis zu 70 Jahre für Tonträger (§ 85 UrhG). Mehrheitlich wird die Leistung zur Generierung von Bild-, Ton- oder Textmaterial geschützt, d. h., die Person, die den Aufnahmeauslöser betätigt, erhält das originäre Verwertungsrecht, das jedoch bei Leistungen, die im Rahmen von Angestellten- oder Auftragsverhältnissen erbracht werden, auf vertraglicher Basis ausdrücklich oder stillschweigend der arbeit- oder auftraggebenden juristischen Person übertragen wird.[169] Ist eine erhebliche Investition für die Erstellung des Datenträgers (vor allem in Bezug auf Datenbanken, aber auch bei Filmen oder Tonträgern) erforderlich, wird der herstellenden juristischen Person, die diese Investition tätigte, das Verwertungsrecht direkt gewährt. Damit liegt die Verfügungsmacht über die erzeugten Daten bei der juristischen Person, die diese gespeicherten Datenbestände finanziert hat und entsprechende Nutzungsrechte vergeben kann.

169 Im Hochschulbereich gibt es aufgrund der verfassungsrechtlich gewährleisteten Wissenschaftsfreiheit und der Konstellation, dass berufene Professor/innen ihre Aufgaben frei und selbstständig wahrnehmen, Hochschulangestellte jedoch dienstliche Aufgaben erfüllen, einige Besonderheiten bezüglich der vertraglich zu konkretisierenden Übertragung von Nutzungsrechten an die Hochschule (ausführlicher z. B. Kuschel 2018).

Die übertragbaren Verwertungsrechte sollen es schöpferisch tätigen Personen ermöglichen, Art und Umfang der Nutzung ihrer Werke vertraglich zu definieren und mit der Zahlung einer Vergütung zu verbinden. Sie sind in unterschiedlichen kreativen Bereichen ein zentrales Element diverser Geschäftsmodelle und Verwertungsketten, zu denen Verlage aber auch Bibliotheken, Sammlungen, Archive und Repositorien sowie Verwertungsgesellschaften gehören können. Über die Reichweite immaterieller Eigentumsrechte bzw. die Weiterverwendungsmöglichkeiten von schöpferischen Werken durch Dritte wird seit Jahren gerungen. Der Begriff des Leistungsschutzes wird nicht nur im Urheberrecht verwendet. Auch bei einigen Produktarten mit besonderer Kritikalität für Mensch und Umwelt (z. B. Arzneimittel, Chemikalien), deren Sicherheit und Wirksamkeit im Rahmen der Produktentwicklung geprüft werden müssen, werden die dafür nötigen Investitionen temporär geschützt (Kasten 8.1).

Kasten 8.1 Leistungsschutz und Datennutzung im Produktrecht

Für einige zulassungspflichtige, forschungsintensive Produkte definiert das jeweilige Produktrecht einen spezifischen Leistungsschutz in Form von Datenexklusivität und Unterlagenschutz (TAB 2017a, S. 175 ff.). Vorreiter sind das Arzneimittel- und das Chemikalienrecht. Entsprechenden Substanzen werden per se besondere Wirksamkeiten auf Menschen, Tiere und/oder Umwelt unterstellt. Deshalb gilt im Umgang mit ihnen ein grundsätzliches Verbotsprinzip mit Erlaubnisvorbehalt. Um eine Marktzulassung zu erhalten, müssen diverse Studien durchgeführt und Daten erhoben werden, anhand derer die substanzspezifische Sicherheit (Schädigungspotenzial, Umweltverträglichkeit) und Wirksamkeit (Nutzenpotenzial) belegt werden. Wenn bei diesen Studien Tests an Menschen vorgenommen werden, sind auch diese genehmigungspflichtig. Dafür müssen studienbeantragende Stellen zum einen darlegen, dass der erwartete Erkenntnisgewinn gegenüber den mit der Studie einhergehenden Sicherheitsrisiken überwiegt – die Studie somit ethisch vertretbar ist. Zum anderen müssen sie jede klinische Studie mittels definierter Metadatensätze (u. a. verwendete Substanz, Studienverantwortlichkeit, Studiendesign, Kurzbeschreibung) registrieren lassen. Sie sind für die Studiendurchführung, die Richtigkeit und den Schutz der erhobenen Daten verantwortlich und haften bei Schäden. Sie sollen zudem zumindest Zusammenfassungen der Studienergebnisse, nicht aber die generierten Primärdaten öffentlich zugänglich machen (wissenschaftliche Publikation). Diese Daten und die Studienergebnisse legen sie im Rahmen der Produktzulassung bzw. -zertifizierung den jeweiligen Prüfinstanzen vor, die diese als Geschäftsgeheimnisse behandeln. Um die jeweiligen Investitionen für die Produktentwicklung zu schützen, werden neben Patenten befristete exklusive

gewerbliche Nutzungsrechte an den Primärdaten und den Zulassungsunterlagen gewährt (in Europa in der Regel 10 Jahre). Erst danach erhalten Hersteller wirkstoffgleicher Substanzen (Generika) eine Marktzulassung.

Im Arzneimittelbereich müssen die Primärdaten auch nach Ablauf der Schutzfristen nicht zugänglich gemacht werden, zumal es sich regelmäßig um personenbezogene Daten besonderer Kategorie handelt (Kasten 3.6). Es gibt jedoch ein vereinfachtes Zulassungsverfahren nach Ablauf der Schutzfrist. Konkurrenten müssen lediglich nachweisen, dass ihr Generikum weitgehend identisch zum Referenzarzneimittel ist und können in Bezug auf Sicherheit und Wirksamkeit auf die Zulassungsunterlagen und Primärdaten des Originals verwiesen (§ 24b AMG). Im Chemikalienrecht ist nach der Sperrfrist eine gemeinsame Datennutzung vorgesehen, wobei es sich bei den primären Studiendaten nicht um personenbezogene Daten besonderer Kategorie handelt und sekundärnutzende Stellen Ausgleichszahlungen an Dateneigner entrichten müssen (Art. 62 ff. der Verordnung [EU] 528/2012[170]).

Seit Jahren wird über den Zugang zu klinischen Studiendaten diskutiert (Rüschemeyer 2020; Taichman et al. 2016). Verfahren für einen gesicherten Datenzugang gibt es bisher nicht. Dadurch können Data-Mining-Aktivitäten Dritter vollständig verhindert werden. Der Ansatz im Chemikalienrecht, der den Aufwand für die Datenerhebung auf mehrere Schultern verteilt, ermöglicht eine Datenweiterverwendung durch Dritte tendenziell eher.

Beschränkung immaterieller Eigentumsrechte und Ausweitung der Data-Mining-Möglichkeiten

Ebenso wie das (immaterielle) Eigentumsrecht (Art. 14 GG, Art. 17 GRCh[171]) gehört die Freiheit, wissenschaftlich zu arbeiten und zu forschen in Deutschland und Europa zu den bürgerlichen Grundrechten (Art. 5 GG, Art. 13 GRCh). Kollidieren Grundrechte ist zwischen diesen abzuwägen und ein gesetzlicher Ausgleich herzustellen. Vervielfältigungen von größeren Teilen geschützter Werke sind in vielen Ländern für nichtkommerzielle Forschungszwecke ohne Autorisierung zulässig. In Deutschland sind derartige Vervielfältigungen traditionell mit einem Vergütungsanspruch urhebender Personen verknüpft, der kollektiv mittels Verwertungsgesellschaften und Reproduktionspauschalen realisiert

170 Verordnung (EU) Nr. 528/2012 über die Bereitstellung auf dem Markt und die Verwendung von Biozidprodukten Text von Bedeutung für den EWR

171 Charta der Grundrechte der Europäischen Union (2000 C 364/01)

werden kann.[172] Da durch die Digitalisierung neue Nutzungsformen urheberrechtlich geschützter Werke möglich werden, wird die Grundrechteabwägung und eine Angleichung bestehender Regelungen immer wieder thematisiert.

2018 wurde das nationale Urheberrecht an die Erfordernisse der Wissensgesellschaft angeglichen und die in Kombination mit Vergütungspauschalen zulässigen nichtgewerblichen Nutzungsbefugnisse ausgeweitet.[173] Zum einen wurde der Umfang der zulässigen Vervielfältigung geschützter Werke zu Bildungs-, Archivierungs- und Forschungszwecken ausgeweitet. Zum anderen wurde Text- und Data-Mining zu nichtkommerziellen Forschungszwecken im Urhebergesetz verankert. Dazu darf eine Vielzahl urheberrechtlich geschützter Werke automatisiert vervielfältigt werden (Ursprungsmaterial), um einen maschinell analysierbaren Korpus zu erstellen, der wiederum einem begrenzten Personenkreis für die gemeinsame wissenschaftliche Forschung sowie Dritten zur Prüfung der Resultate zugänglich gemacht werden darf. Um die Vorgaben der guten wissenschaftlichen Arbeit einhalten zu können, dürfen Vervielfältigungen des Ursprungmaterials und der erstellte Korpus langfristig in Bibliotheken und Forschungseinrichtungen archiviert werden (§ 60d Abs. 3 UrhG).

Treiber für die Aufnahme von Text- und Data-Mining ins Urheberrecht waren vor allem die Entwicklungen bei digitalen Literaturdatenbanken und deren Recherchewerkzeuge zur Nutzung wissenschaftlicher Publikationen (ausführlicher z. B. Schüller-Zwierlein/Leiwesmayer 2018, S. 25 ff.). Diese Urheberrechtsreform wurde kontrovers diskutiert. Mehr als 100 Stellungnahmen gingen ein.[174] Insbesondere wissenschaftsnahe Institutionen begrüßten die nutzerfreundlich ausgestalteten Schrankenbestimmungen und die Verankerung von Text- und Data-Mining im Urheberrecht. Kritik wurde vor allem von wirtschaftlich agierenden Organisationen geäußert. Unter anderem wurde auf die Unschärfe etlicher Begriffe hingewiesen, der Zeitpunkt als verfrüht bezeichnet (den diesbezüglichen europäischen Aktivitäten solle nicht vorgegriffen werden) und die Annahme bezweifelt, dass die Regelungen keinen oder kaum Einfluss auf die jeweiligen Marktstrukturen hätten. Aus der datenanalytischen Perspektive ergaben sich Herausforderungen u. a. zur Reichweite etlicher Begriffe, zum Auf- und Ausbau von Datenrepositorien und -infrastrukturen, die einen regelkonformen Datenzugang ermöglichen sollen, sowie zu nachhaltigen Geschäftsmodellen, einschließlich Kosten- und Erlösbeteiligungen.

»Wissenschaftliche Forschung« und »nichtkommerzielle Zwecke« sind Kernbegriffe, deren Interpretation war und ist schwierig: Wie weit reichen diese

172 In etlichen anderen Ländern, wie z. B. den USA, Großbritannien, Israel oder Südkorea gibt es keinen Vergütungsanspruch bei nichtkommerzieller Nutzung zu Bildungs- und Forschungszwecken (definiert durch Fair-Use-Klauseln im Copyright).

173 Gesetz zur Angleichung des Urheberrechts an die aktuellen Erfordernisse der Wissensgesellschaft (UrhWissG)

174 www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/DE/UrhWissG.html (2.11.2021)

Begriffe? Wo hören wissenschaftliche Forschung und nichtkommerzielle Datennutzung auf und wo fängt die gewerbliche Entwicklung von digitalen Produkten und Diensten an? Wo sind Citizen-Science-Ansätze zu verorten? Ist eine wissenschaftliche Forschung auch noch mit nichtkommerziellen Zwecken verknüpft, wenn ein Prototyp eines Modells oder Algorithmus für die Finalisierung und Einbettung in Anwendungssoftware in einem ausgegründeten Spin-off finalisiert oder an gewerbliche Entwickler veräußert wird? Wie sind Algorithmen zu bewerten, die perspektivisch u. a. in komplexere datenanalytische Software integriert oder auf digitalen Plattformen angeboten werden sollen und in gestuften Versionen kostenlose, werbefinanzierte oder kostenpflichtige Dienstleistungen ermöglichen? Wer entscheidet bei derartigen Fragen? Wie lassen sich die Vorgaben überwachen? Dürfen ausschließlich öffentliche oder akademische Einrichtungen bzw. öffentlich finanzierte Forschungsprojekte derartig privilegiert werden? Können die Potenziale von Text- und Data-Mining-Ansätzen ausgeschöpft werden, wenn der Digitalwirtschaft die Datennutzung verwehrt wird?

2018 wurde auch festgelegt, dass die Regelungen nach 4 Jahren evaluiert werden, um dann über deren Fortbestand zu entscheiden (§ 142 UrhG).[175] Die nationale Urheberrechtsreform von 2018 griff den europäischen Aktivitäten vor. Die Richtlinie (EU) 2019/790 über das Urheberrecht und die verwandten Schutzrechte im digitalen Binnenmarkt (DSM-RL) enthielten ebenfalls Formulierungen zum Text- und Data-Mining, die eine weitere Anpassung des nationalen Urheberrechts bis Mitte 2021 erforderten, ohne dass die Ergebnisse der vereinbarten Evaluation vorlagen und berücksichtigt werden konnten. Mit dem Gesetz zur Anpassung des Urheberrechts an die Erfordernisse des digitalen Binnenmarktes wurden die europäischen Vorgaben 2021 in nationales Recht umgesetzt.

- Zum einen wurden die Regelungen zum Text- und Data-Mining für wissenschaftliche Forschungszwecke angepasst: berechtigte Forschungsorganisationen und nichtkommerzielle Forschungszwecke wurden genauer definiert (§ 60d UrhG) und die pauschalierte Vergütung gestrichen (§ 60h Abs. 2 Nr. 3 UrhG).
- Zum anderen wurden Text- und Data-Mining allgemein zugelassen, sofern Rechtsinhabende dem nicht widersprochen haben (Opt-out-Ansatz). Vervielfältigungen von digitalen Werken sind dafür zulässig und müssen im Anschluss gelöscht werden (§ 44b UrhG).

175 Der Evaluationsbericht soll dem Bundestag bis zum 1. März 2022 vorgelegt werden. Dazu erstellte Stellungnahmen unter www.bmj.de/SharedDocs/Gesetzgebungsverfahren/DE/UrhWissG_Evaluation.html;jsessionid=6869C726ED2EF3ECC400C13B3CA28E68.2_cid324?nn=6712350 (4.5.2022)

Auch im Rahmen dieses Gesetzgebungsprozesses gaben unterschiedliche Institutionen Stellungnahmen ab.[176] Umstritten sind die unterschiedlichen Verfahren für nichtkommerzielle Forschungszwecke und andere Absichten, die trotz begrifflicher Konkretisierung praxisfern seien, weil Data-Mining eine hochrelevante Technik der Digitalwirtschaft sei und man den Prozess im weiteren Sinne betrachten müsse, der vielfältige Algorithmen und Digitalprodukte hervorbringe, die kommerziell weiterverwendet würden. Umstritten ist auch die teilweise Vergütungsfreiheit. Einerseits ist die Verteilung der Vergütung aufwendig und kompliziert, zumal einzelne Werke bei Big-Data-Ansätzen nur einen minimalen Beitrag leisten und Data-Mining-Weiterverwendungen die primären Geschäftsmodelle nicht beeinträchtigen würden. Andererseits wird nicht berücksichtigt, dass viele kreativ Tätige, die Teile des Ursprungmaterials für Data-Mining schaffen, von ihrer Tätigkeit leben müssen und deshalb auch am sekundär erzeugten Mehrwert beteiligt werden wollen. Zudem wird darauf hingewiesen, dass die im Rahmen der nichtkommerziellen Forschung zulässige Archivierung von Ursprungsmaterial und Korpora zur Prüfung von Forschungsergebnissen aufwendig ist und früher oder später Fragen zur Sekundärnutzung aufwerfen wird.

Mit der Urheberrechtsreform von 2018 hat der Deutsche Bundestag den Gesetzgeber aufgefordert, die 2018 eingeführten Regelungen nach 4 Jahren zu evaluieren. Dazu hat zum einen das Bundesministerium für Justiz vielfältige Akteure um Stellungnahmen gebeten und im Namen der Bundesregierung einen Evaluierungsbericht erstellt (Bundesregierung 2022). Ergänzend hat das Bundesministerium für Bildung und Forschung dazu eine Studie mit qualitativen Interviews in Auftrag gegeben, um die Sicht von Praktiker/innen zusätzlich einzuholen (Kreutzer/Fischer 2022). Im Rahmen der Evaluation wurden die gesetzlichen Regelungen zum Text- und Data-Mining grundsätzlich als praktikabel bewertet. Die Abgrenzung zwischen nichtkommerzieller und kommerzieller Forschung sei in der Praxis oftmals schwierig. Rechtsunsicherheiten gebe es bezüglich des Personenkreises und der Dauer der Bereitstellung vervielfältigter Werke und erstellter Korpora, der Aufbewahrungsfristen sowie den Möglichkeiten und Grenzen der Nachnutzung dauerhaft gespeicherter Korpora (Bundesregierung 2022, S. 12). Unterschiedliche Positionen gab es zur praktischen Relevanz der Regeln zum Text- und Data-Mining. Einerseits wird auf die wachsende Bedeutung entsprechender Ansätze hingewiesen, andererseits gebe es bisher nur eine moderate Nutzung geschützter Werke zum Text- und Data-Mining auf Basis von § 60d UrhG (Bundesregierung 2022, S. 56).

176 www.bmj.de/SharedDocs/Gesetzgebungsverfahren/DE/Gesetz_Anpassung-Urheberrecht-dig-Binnenmarkt.html (4.5.2022)

Grenzen des Urheberrechts

Die Regularien des Urheberrechts zum Text- und Data-Mining greifen nur bei Datenwerken und -beständen, die öffentlich zugänglich sind. Wissenschaftliche Bibliotheken als ein Treiber dieser Entwicklung halten in ihren Repositorien bisher vor allem wissenschaftliche Texte, in denen ein methodisches Vorgehen skizziert und erzielte Ergebnisse diskutiert werden, nicht aber die diesen Texten zugrundeliegenden Analysedatensätze. Aus der datenanalytischen Perspektive stellt sich vor allem die Frage nach der Zugänglichkeit zu expliziten Datenrepositorien und dem Auf- und Ausbau von Dateninfrastrukturen, die diesen Zugang rechtssicher gewährleisten sollen.

Die Tatsache, dass Datenwerke und -bestände trotz Urheber- und Leistungsschutzrechten für Data-Mining zu wissenschaftlichen Forschungszwecken verwendet werden dürfen, sollte nicht darüber hinwegtäuschen, dass vielfältige, zunehmend maschinell generierte (Roh-)Datenbestände aufgrund der zu geringen Schöpfungs-, Leistungs- oder Investitionshöhe gar nicht vom Urheber- oder Leistungsschutzrecht erfasst werden. An derartigen Datenbeständen gibt es zwar formaljuristisch keine immateriellen Eigentumsrechte, jedoch ist die Stelle, die die Verfügungsgewalt über die datenerhebenden und -speichernden Medien hat, de facto im Besitz dieser Daten. Oftmals sind sie per se unzugänglich, weil die jeweilige datenverarbeitende Stelle Rechte auf Geschäftsgeheimnisse hat, weil es keine praktikablen Datenzugangsstrukturen gibt und/oder weil die Daten aufgrund ihrer Kritikalität auf gesetzlicher Grundlage geschützt werden müssen (z. B. DSGVO oder SatDSiG). Die datenverarbeitende Stelle hat dadurch eine faktische Nutzungsexklusivität, auch wenn ihr urheberrechtlich keine exklusiven Verwertungsrechte zuerkannt werden.

In der Auseinandersetzung um die rechtliche Weiterentwicklung des Umgangs mit Daten wird mitunter der Begriff des Dateneigentums diskutiert. In der rechtswissenschaftlichen Diskussion überwiegt derzeit die Skepsis gegenüber einem Dateneigentum. Expertengremien empfehlen immaterielle Eigentumsrechte an Daten nicht über das bestehende Maß hinaus zu vertiefen (DEK 2019, S. 18; Enquete-Kommission 2020, S. 183). Bei personenbezogenen Daten würden damit eher neue Probleme entstehen. Wichtiger wäre es, den rechtskonformen Zugang zu Daten(beständen) und die Datennutzung stärker in den Blick zu nehmen.

8.2 Abbildungen

Abb. 2.1 Data-Mining: schematische Darstellung der Prozessschritte 38
Abb. 2.2 Position der Choleraopfer und der Wasserbrunnen in London 1854 40
Abb. 2.3 Datenstrukturen (schematische Darstellung) 43
Abb. 2.4 Grobarchitektur von Datenhaltung und Analyse 48
Abb. 2.5 Teilung eines Analysedatenbestandes zur Kreuzvalidierung 63
Abb. 3.1 Verschlüsselung, Pseudonymisierung, Anonymisierung und Vergröberung von Datensätzen 110
Abb. 4.1 Erhebung medizinischer Datenbestände und deren Zugänglichkeit im Überblick 160
Abb. 4.2 Radiologische Befundung von Mammografien 173
Abb. 4.3 Schematische Darstellung prognostischer Multigentests 181
Abb. 4.4 Leistungsabrechnungsdaten der gesetzlichen Krankenversicherung 200
Abb. 5.1 Regionale MRSA-Diagnosen in Deutschland (2013) 224
Abb. 5.2 Datenmodell: nationaler Versorgungsdatenbestand 246
Abb. 5.3 Vierfeldertafel zum Finden unerwünschter Arzneimittelwirkungen 257

8.3 Tabellen

Tab. 2.1 Statistische Gütekriterien von Klassifikationsverfahren 65
Tab. 3.1 Kritikalitätsstufen algorithmischer Systeme 134
Tab. 4.1 Programme zur Unterstützung med. Entscheidungen (Auswahl) 188

8.4 Kästen

Kasten 2.1 Codierungen von Objekten und Merkmalen (DIN 6763) 45
Kasten 2.2 GAIA-X 52
Kasten 2.3 Beispielhafte externe Prüfung voreingestellter Analysetools 67
Kasten 3.1 Kritische Infrastrukturen 78
Kasten 3.2 Europäische Regulierungsinitiativen zur Verbesserung der Datennutzung 81
Kasten 3.3 Open-Data-Konzepte 82
Kasten 3.4 Amtliche Geoinformationssysteme ATKIS und ALKIS 86
Kasten 3.5 Rechtsgrundlagen der Geodateninfrastruktur 91
Kasten 3.6 Personenbezogene Daten besonderer Kategorie 102
Kasten 3.7 Reidentifizierungsexperiment 113
Kasten 3.8 Einwilligungsmodelle 121

Kasten 4.1 Ärztliche Schweigepflicht und medizinische Daten 139
Kasten 4.2 Standardisierung und Interoperabilität in der Medizin 147
Kasten 4.3 Codierungen und Klassifikationen zur Leistungs- abrechnung 197
Kasten 5.1 Data-Mining im Rahmen der Pharmakovigilanz (Beispiele) 252
Kasten 8.1 Leistungsschutz und Datennutzung im Produktrecht 297

8.5 Abkürzungen

ABDA	Bundesvereinigung Deutscher Apothekerverbände
AGB	Allgemeine Geschäftsbedingungen
AOK	Allgemeine Ortskrankenkasse
ARZ	Apothekenrechenzentren
BAN	bundeseinheitliche Arztnummer (von den Ärztekammern an alle approbierten Ärzt/innen vergeben)
BAS	Bundesamt für Soziale Sicherung
BDSG	Bundesdatenschutzgesetz
BfArM	Bundesinstitut für Arzneimittel und Medizinprodukte
BGB	Bürgerliches Gesetzbuch
BMBF	Bundesministerium für Bildung und Forschung
BMG	Bundesgesundheitsministerium
BMV-Ä	Bundesmantelvertrag Ärzte
BMWi	Bundesministerium für Wirtschaft und Energie
BSI	Bundesamt für Sicherheit in der Informationstechnik
BSNR	Betriebsstättennummer (von den Kassenärztlichen Vereinigungen für Arztpraxen vergeben, die Leistungen zu Lasten der GKV erbringen)
BStatG	Bundesstatistikgesetzes
CAD	computer-assisted detection (Software medizinischen Bildbefundung)
DAPI	Deutsche Arzneiprüfungsinstitut
DaTraV	Datentransparenzverordnung
DEK	Datenethikkommission
DFG	Deutsche Forschungsgemeinschaft
DGU	Deutsche Gesellschaft für Unfallheilkunde
DiGA	digitale Gesundheitsanwendungen
DIMDI	Deutsches Institut für Medizinische Dokumentation und Information (seit 2020 Teil des BrArM)
DIN	Deutsches Institut für Normung
DRG	Diagnosis Related Groups (diagnoseorientierte Fallgruppe)
DSFA	Datenschutz-Folgenabschätzung
DSGVO	Datenschutz-Grundverordnung
DVG	Digitale-Versorgung-Gesetz
EBM	Einheitlicher Bewertungsmaßstab
FDA	Food and Drug Administration (US-amerikanische Lebensmittel und Arzneimittel-Agentur)
FDI	Forschungsdateninfrastruktur
G-BA	Gemeinsamer Bundesausschusses

GDI-DE	Geodateninfrastruktur Deutschland
G-DRG	German Diagnosis Related Groups (nationale diagnoseassoziierte Fallpauschalen zur Vergütung stationärer Behandlungsleistungen)
GeoZG	Geodatenzugangsgesetz
GG	Grundgesetz
GIS	Geoinformationssystem
GKK	gesetzliche Krankenkasse
GKV	gesetzliche Krankenversicherung
HPO	Human Phenotype Ontology (eine an der Charité federführend entwickelte Ontologie zur Beschreibung menschlicher Phänotypen)
ICD	International Statistical Classification of Diseases and Related Health Problems (Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme)
ICGC	International Cancer Genome Consortium
ID	Identifikations(nummer)
IK	Institutionenkennzeichen (von der Arbeitsgemeinschaft Institutionenkennzeichen vergeben für Einrichtungen zu Leistung Lasten jeglicher Sozialleistungsträger erbringen
InEK	Institut für das Entgeltsystem im Krankenhaus
IT	Informationstechnologie
KBV	Kassenärztliche Bundesvereinigung
KI	künstliche Intelligenz (Artificial Intelligence – AI)
KNN	künstliche neuronale Netze
KV-Nr.	Krankenversicherungsnummer
KRITIS	Kritische Infrastrukturen
KVen	Kassenärztliche Vereinigung(en)
LANR	lebenslange Arztnummer (von den Kassenärztlichen Vereinigungen für niedergelassene Ärzt/innen vergeben, die Leistungen zu Lasten der GKV erbringen)
MBO-Ä	Musterberufsordnung der in Deutschland tätigen Ärztinnen und Ärzte
Morbi-RSA	morbiditätsorientierter Risikostrukturausgleich
mRNA	messenger ribonucleic acid (Boten-Ribonukleinsäure)
MRSA	Methicillin-resistenter Staphylococcus aureus (auch multiresistente Keime)
MSKCC	Memorial Sloan Kettering Cancer Center
NFDI	Nationale Forschungsdateninfrastruktur
NCT	Nationales Centrum für Tumorerkrankungen (Heidelberg)
PIS/KIS	Praxis- oder Krankenhausinformationssysteme
PKV	private Krankenversicherung
PPN	Pharmacy Product Number (europäisches Nummernsystem für Arzneimittel)
ProdHaftG	Produkthaftungsgesetz
PZN	Pharmazentralnummer (nationales Nummernsystem für Arzneimittel)
QS	Qualitätssicherung
RatSWD	Rat für Sozial- und Wirtschaftsdaten

RKI	Robert Koch Institut
SatDSiG	Satellitendatensicherheitsgesetz
SGB	Sozialgesetzbuch
SNOMED	Systematized Nomenclature of Medicine
UAW	unerwünschte Arzneimittelwirkungen
UrhG	Urheberrechtsgesetz
WIdO	wissenschaftliches Institut der Ortskrankenkassen
Zi	Zentralinstitut für die Kassenärztliche Versorgung

BÜRO FÜR TECHNIKFOLGEN-ABSCHÄTZUNG BEIM DEUTSCHEN BUNDESTAG

Karlsruher Institut für Technologie

Neue Schönhauser Straße 10
10178 Berlin

Telefon: +49 30 28491-0
E-Mail: buero@tab-beim-bundestag.de
Web: www.tab-beim-bundestag.de
Twitter: @TABundestag