

Data-Mining – Imagewandel der Statistik?

Dem Umgang mit Daten haftete lange Zeit kein besonders gutes Image an: Die genaue zahlenmäßige Erfassung vielfältiger Sachverhalte insbesondere durch staatliche Einrichtungen wurde teils als Erbsenzählerei titulierte, teils als unzulässiger Übergriff massiv infrage gestellt. Und dass man nur den Statistiken trauen sollte, die man selbst gefälscht hatte, ist ein gängiges Bonmot. Dieses abwertende Bild zum Umgang mit Daten wandelt sich derzeit zumindest in Teilen der Gesellschaft erheblich – wobei die Einstellungen zwischen hohen Nutzungserwartungen sowie alter und neuer Skepsis mäandern. Einerseits wird eine unglaubliche Dateneuphorie geschürt. Neue Begriffe werden etabliert, die positiv konnotiert sein und spritzig klingen sollen, semantische Unschärfen lassen bewusst Spielraum für Interpretation: Daten werden als Rohstoffe der Wissensgesellschaft oder als Öl des 21. Jahrhunderts bezeichnet, wobei die Bestände immer größer (Big Data) und die Techniken zu deren Erhebung, Verwaltung und Analyse als immer smarter oder intelligenter bezeichnet werden. Dazu passt der Begriff »Data-Mining« (wörtlich »Datenbergbau«), der das suggestive Bild des Schürfens nach Rohstoffadern, der Suche nach Nuggets im heutigen Datenzeitalter heraufbeschwört. Andererseits misstraut man der um sich greifenden Datenerfassung und deren potenziellen Folgen. Auch hier werden Metaphern bemüht wie Datenkraken oder das Ende der Privatheit.

Die Skepsis wird u. a. dadurch befördert, dass sich zumindest Teile der Gesellschaft weder als Herr über ihre Daten sehen noch die Analysetechniken verstehen und deren Reichweite abschätzen können. Datenbestände erscheinen kaum sicherbar, Grenzziehungen zwischen zulässiger Datennutzung und Datenmissbrauch als Ansichtssache und die Überwachung der Regelinhaltung kaum realisierbar.

Vor diesem Hintergrund wurde das TAB 2014 mit dem Projekt »Data-Mining – gesellschaftspolitische und rechtliche Herausforderungen« beauftragt. Anhand von zwei Anwendungsbereichen, der Nutzung von Geodaten einerseits und von personenbezogenen Gesundheitsdaten andererseits, sollen mit Data-Mining-Verfahren einhergehende gesellschaftsrelevante Fragestellungen dargestellt und diskutiert werden. Derzeit wird im TAB an der Zusammenführung der Ergebnisse zweier großer Gutachten, der Prüfung bzw. Interpretation der Aussagen und der Fertigstellung des Arbeitsberichts gearbeitet. Ohne die Ergebnisse des Projekts vorwegnehmen zu wollen, soll nachfolgend eine Annäherung an den Begriff Data-Mining gemacht und an zwei Beispielen aus dem

Gesundheitsbereich veranschaulicht werden, welche Arbeitsschritte und welcher Aufwand mit Data-Mining-Verfahren verbunden sind. Die Nutzung elektronischer Gesundheitsdaten zu Forschungszwecken wird auf nationalstaatlicher Ebene sehr unterschiedlich gehandhabt. Daher wird abschließend die Situation in Deutschland mit der in anderen Ländern kurz verglichen.

Was ist Data-Mining?

Data-Mining im engen Sinn meint die Anwendung von (komplexen) mathematisch-statistischen Verfahren und Algorithmen, um in großen Datenbeständen Regelmäßigkeiten, Zusammenhänge und (neue) Muster zu erkennen. In diesem Verständnis ist es nur ein Element in einem Prozess, den Fayyad et al. (1996) als »knowledge discovery in databases« bezeichneten und der häufig auch unter dem noch breiter gefassten Schlagwort Big Data verortet wird. Die Begriffe Data-Mining, »knowledge discovery in databases« oder Big Data werden nahezu synonym benutzt, etablierte deutsche Begriffsentsprechungen gibt es nicht. Im Kern zielen alle Ver-

fahren darauf ab, aus großen, nur noch maschinell verwalt- und verarbeitbaren Datenmengen möglichst schnell (neue) Muster zu erkennen und Wissen zu generieren. Eine kategoriale Abgrenzung der dafür genutzten Verfahren erscheint schwierig. Allgemein werden unter Data-Mining komplexe und aufwendige Auswertungsverfahren gefasst und damit einerseits von trivial-einfachen statistischen Standardverfahren (z. B. die grafische Darstellung von [Index-]Zahlen im Zeitverlauf) und andererseits von vollautomatisierten Prozesssteuerungen (z. B. zunehmend autonom arbeitende Insulinpumpen) abgegrenzt. Durch die Abgrenzung zu automatisierten Prozessen wird deutlich, dass bei Data-Mining-Verfahren menschliche Entscheidungen nach wie vor relevant sind. Die Abgrenzung ist jedoch dynamisch; was vor Jahren als komplex und aufwendig galt, kann durch den technischen Fortschritt trivial-einfach werden, und Algorithmen, die zunächst lediglich einzelne Arbeitsschritte autonom durchführen, können im Laufe der Entwicklung immer umfangreichere und komplexere Prozesse automatisiert bewältigen (Bernsdorf et al. 2015, S. 36; Schepers et al. 2015, S. 20).

Insbesondere wenn der Frage nach den mit den Datenanalysetechniken verbundenen gesellschaftlichen Chancen und Herausforderungen nachgegangen werden soll, liegt es auf der Hand, nicht nur die Anwendung komplexer Verfahren und Algorithmen zur Erkennung (neuer) Muster in Datenbeständen (Data-Mining im engen Sinn), sondern den gesamten Prozess der Wissensgewinnung aus Datenbeständen zu betrachten (Data-Mining im weiteren Sinn). Dieser Prozess kann durch folgende Schritte strukturiert werden (Fayyad et al. 1996):

- Festlegung der Aufgabenstellung: Je nach Zielstellung ist grundsätzlich zu klären, ob z. B. ein Prognosemodell erstellt, ein Zusammenhang zwischen Ereignissen gefunden oder ein Objekt/Vorgang klassifiziert werden soll.

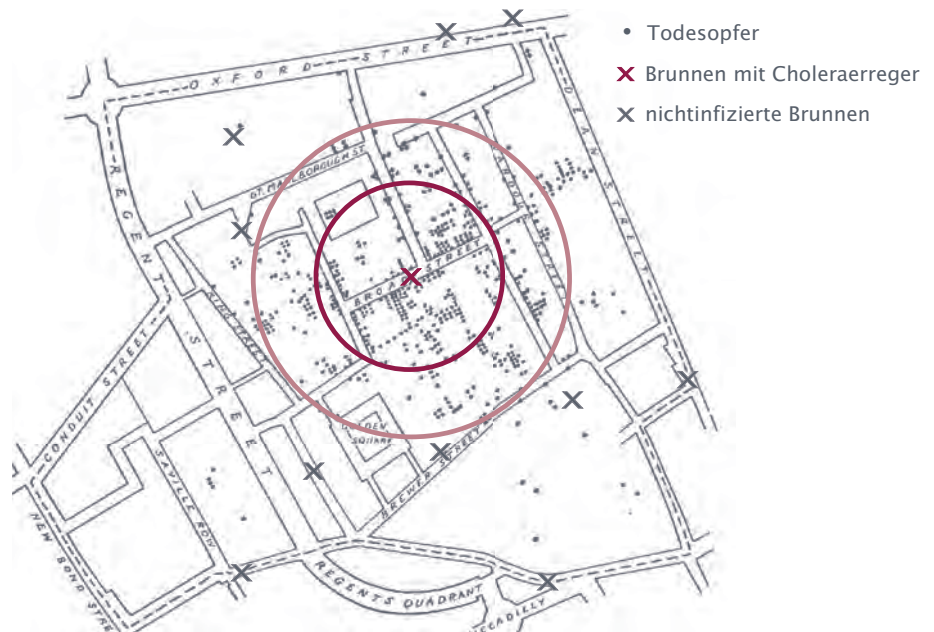
- > Datenaufbereitung: Je nach Aufgabe sind relevante Datenbestände auszuwählen (dabei u. a. Zugriffsrechte zu klären), zu bereinigen (z. B. Umgang mit fehlenden und fehlerhaften Werten klären) und zu transformieren (z. B. Variablen umrechnen oder zusammenfassen).
- > Datenanalyse: Je nach Aufgabenstellung und Datentypen (Zahlen, Orts-/Zeitangaben, Zeichenketten/Texte, Bilder) kann mit spezifischen Verfahren nach Mustern, Strukturen und Besonderheiten gesucht werden.
- > Bewertung/Interpretation der Ergebnisse: Analyseergebnisse können intern und extern bewertet werden. Interne Evaluierungen stützen sich meist auf mathematisch-statistische Gütekriterien (z. B. Signifikanz, Sensitivität, Effektstärke, Bestimmtheitsmaße) und geben Hinweise auf die Qualität der aus den Daten abgeleiteten Ergebnisse. Für die Prüfung der fachlichen Richtigkeit/Wahrheit sind externe Evaluierungen erforderlich (fachliche Diskussionen, Abgleich mit anderen fachwissenschaftlichen Ergebnissen, um die Richtigkeit gefundener Aussagen zu untermauern). Diese externen Evaluationen gehen über den Data-Mining-Prozess hinaus. Sie sind zur Ergebnisbewertung im Rahmen der eigentlichen Aufgabenstellung jedoch wichtig.

Anhand eines historischen und eines heutigen Anwendungsbeispiels sollen diese Prozessschritte veranschaulicht werden.

Historisches Beispiel der Wissensgenerierung aus Daten

Um 1850 gab es zu unterschiedlichen Krankheiten diverse Annahmen und Theorien, (z. B. dass sie durch üble Dünste oder Kontakt zu Kranken übertragen werden), abgesichertes Wissen zu Infektionskrankheiten, deren mikrobiologischen Ursachen und Verbreitungswe-

Abb. 1 Position der Choleraopfer und der Wasserbrunnen in London 1854



Quelle: Bernsdorf et al. 2015, S. 47, nach [https://de.wikipedia.org/wiki/John_Snow_\(Arzt\)](https://de.wikipedia.org/wiki/John_Snow_(Arzt))

gen waren jedoch begrenzt. Dennoch vermuteten einige Mediziner, dass das Übertragungsmedium von Cholera keine Dünste, sondern vielmehr Wasser sei. Einer von ihnen war der Arzt John Snow, der 1854, als im Londoner Stadtteil Soho eine Choleraepidemie ausbrach, eine Karte dieses Stadtteils zeichnete und darin den Wohnort jedes Choleraopfers markierte. Da er die These vertrat, dass verunreinigtes Wasser die Ursache der Epidemie war, markierte er in dieser Karte auch die Positionen der örtlichen Wasserbrunnen (Abb. 1).

Allein durch die grafische Darstellung wurde die Häufung der Todesopfer um ein Zentrum offensichtlich. In diesem Zentrum befand sich der Brunnen in der Broad Street. Snow interpretierte sein Analyseergebnis nicht nur als Verbindung zweier Sachverhalte (Wasserbrunnen und Choleraopfer), sondern unterstellte eine Ursache-Wirkungs-Beziehung. Obwohl auch weiter entfernt einige Todesfälle zu verzeichnen waren und er zu diesem Zeitpunkt keine biologisch dezidierte Begründung liefern

konnte, forderte er die sofortige Stilllegung des Brunnens. Wahrscheinlich stieß er nicht allseits auf offene Ohren, widersprach er doch damals vorherrschenden Meinungen zur Krankheitsübertragung. Überliefert ist, dass er eigenmächtig den Pumpenschwengel in der Broad Street abmontiert habe, woraufhin die Zahl der Choleraopfer sank (Gerste 2014).

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 (Adressen) hergestellt. Was er 1854 mit Zettel und Stift manuell vollzog, wird inzwischen mithilfe von Datenbanken und Analysesoftware bzw. Informationssystemen (die Datenbanken und Analysesoftware vereinen) durchgeführt. Snows Vorgehen bezeichnet man heute als räumliches Clustern und Hotspotanalyse. Die insbesondere 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).

Snow ging nach seiner Interpretation, dass der Brunnen die Ursache für die Cholerafälle sei, analytisch einen Schritt weiter. Um seine von der damaligen Mehrheitsmeinung abweichende Hypothese zu untermauern, verknüpfte er sein Ergebnis aus der Datenanalyse mit weiteren Fachkenntnissen und führte zu zusätzlichen Untersuchungen durch (externe Prüfung). Mit einem Kollegen schaute er sich unter dem Mikroskop sowohl Wasserproben des stillgelegten Brunnens und der Themse (aus der der Brunnen gespeist wurde und in die die Abwasserkanäle mündeten) als auch Stuhlproben der Erkrankten an. Sie fanden überall gleich aussehende kommaförmige Mikroorganismen – die gleichen, die ein italienischer Arzt kurz vorher als *vibrio cholerae* beschrieb (Gerste 2014). Auch diese Entdeckung widersprach insbesondere der Dunstübertragungsthese, auch wenn sie die kausale Beziehung nicht zweifelsfrei belegen konnte. Weitere Belege, dass Cholera durch diesen wasserbezogenen Erreger verursacht wird, lieferten Robert Koch und Kollegen erst drei Jahrzehnte später, als es ihnen gelang, die Erreger zu züchten und den Ansteckungsmechanismus aufzudecken.

Rückblickend fällt das Vorgehen von Snow in die Phase, in der die empirische (datengestützte) Forschung auch in der Medizin an Bedeutung gewann. Dort bedeutet evidenzbasiert inzwischen wesentlich auch datenbasiert. Das Beispiel zeigt, dass das Vorgehen der datengestützten Informations- und Wissensgenerierung lange etablierte Praxis ist. Und auch die eingangs skizzierten Data-Mining-Prozessschritte entsprechen nach wie vor weitgehend denen, die seit Jahrzehnten in Statistikkursen vermittelt werden. Ist Data-Mining also nur ein smarterer Begriff für lange etablierte Verfahren der analytischen Statistik?

Was ist das Neue bei heutigen Data-Mining-Verfahren?

Auch heute werden mit Clusteranalysen ähnliche Objekte zu Gruppen zusammengefasst, mit Disproportionalitätsanalysen Merkmalshäufungen gesucht oder mit Regressionsanalysen Zusammenhänge zwischen Variablen ermittelt. Dennoch sind Veränderungen in der Art und Weise der Datenerhebung und der Datenanalytik unübersehbar.

Einerseits wachsen maschinenlesbare Datenbestände überall in ungeahntem Tempo, da kontinuierlich mehr Lebensbereiche immer genauer von diversen Akteuren mit vielfältigen, oft internetbasierten Geschäftsmodellen digital erfasst und gespeichert werden. Die Verarbeitung von sehr großen Datenmengen geht laut Mayer-Schönberger/Cukier (2013, S. 20 f.) mit einer »Big-Data-Revolution« einher, mit der sich das Bedürfnis nach Exaktheit überwinden ließe und eine Unschärfe in der Datensammlung akzeptierbar wäre. Auch würde sie mit einer Abwendung von der jahrtausendealten Suche nach kausalen Zusammenhängen einhergehen, da es oft ausreicht, zu erkennen, dass etwas geschieht, ohne dass bei allem nach einer Ursache gefragt werden müsste.

Trotz des verbrieften Rechts auf informationelle Selbstbestimmung in Deutschland und Europa hegt wohl kaum noch jemand ernsthaft die Hoffnung, Herr über »seine Daten« zu sein, weder in staatlichen noch in kommerziellen Strukturen. Staatliche Strukturen werden spätestens seit den Enthüllungen von Edward Snowden neu bewertet. Bei kommerziellen Strukturen ist von vornherein klar, dass man sich deren digitale Services (von Google bis Facebook) fast ausnahmslos mit der Preisgabe seiner Datenspuren erkauft. Die Möglichkeiten, auf diese Daten zuzugreifen, und die Kompetenzen, aus ihnen Informationen abzuleiten, Nutzen zu ziehen und davon zu profitieren, konzentrieren sich bisher wesentlich auf wenige Akteure.

Andererseits werden durch die kontinuierliche Steigerung von Speicherkapazitäten und Rechenleistung auch die Datenverwaltungs- und -analyseverfahren vielfältiger und eröffnen neue Möglichkeiten. Neben merkmalsbetonten relationalen Datenbanken und klassischen multivariaten Verfahren, bei denen einzelne Datenobjekte tendenziell im Hintergrund bleiben, werden mit objektbasierten Datenbankstrukturen und objektorientierter Programmierung neue Konzepte und Verfahren entwickelt. Diese Objektorientierung erlaubt eine bessere Sichtweise auf den oder das Einzelne mit seinen jeweiligen Eigenschaften. Mischformen aus beiden (sogenannte objektrelationale Datenbanken) können Daten aus diversen Beständen besonders vielfältig verknüpfen. Zu klassischen Computerprogrammen, bei denen der Mensch Rechenalgorithmen (Handlungsanleitungen) entwickelte und Schritt für Schritt Regeln definierte und programmierte, die Computer abarbeiteten, treten neue Verfahren, z. B. auf künstlichen neuronalen Netzen aufbauende, selbstlernende Systeme, die auf der Basis eigener Berechnungen gesteuert werden. Bei großen Datenmengen funktionieren diese selbstlernenden Systeme besonders gut. Sie liefern teilweise auch dann noch zuverlässige Ergebnisse, wenn konventionelle Berechnungsverfahren an ihre Grenzen kommen.

Auch wenn die in der zweiten Hälfte des vergangenen Jahrhunderts geschürten Visionen zur Schaffung künstlicher Intelligenz bis heute als übersteigert gelten, ist die Entwicklung des rechentechnisch Machbaren dennoch kontinuierlich vorangeschritten und lässt die IT-Branche derzeit von kognitiven Maschinen als Meilenstein der Wissensarbeit schwärmen (Bitkom 2015). Befürworter unterstellen diesen neuen Verfahren derzeit besonders große Potenziale, da sie bei komplexen Sachverhalten in kürzester Zeit bessere Ergebnisse liefern würden, als klassische, fest definierte Algorithmen. Kritiker stoßen sich daran, dass die Vorgehensweise dieser Verfahren nicht nachvollziehbar ist,

da sie aus Trainingsdaten und durch Verstärkung lernen, aber nicht für bestimmte Aufgaben fest programmiert werden (mitunter werden sie als Blackbox angesehen).

Die Gutachter wurden im Rahmen des TAB-Projekts aufgefordert, den derzeitigen Einsatz von Data-Mining-Verfahren einerseits mit Geodaten und andererseits mit Gesundheitsdaten anhand von Anwendungsbeispielen zu veranschaulichen. Deutlich wurde, dass in den beiden öffentlichen Aufgabenbereichen der zivilen Gefahrenabwehr und im Gesundheitssystem nach wie vor die Datenaufbereitung erheblichen Aufwand verursacht und zur Datenanalyse mehrheitlich noch klassische Verfahren genutzt werden. Ein aktuelles Anwendungsbeispiel aus dem Gesundheitsbereich soll veranschaulichen, wie Data-Mining gegenwärtig in Deutschland in öffentlichen Aufgabenbereichen realisiert werden kann.

Beispiel: Signalgenerierung zu unerwünschten Arzneimittelwirkungen (UAW)

Die Sicherheitsüberwachung von Arzneimitteln (Pharmakovigilanz) ist in Deutschland durch das Arzneimittelgesetz als kontinuierliche Aufgabe sowohl der jeweiligen Hersteller (im Rahmen der Produkthaftung) als auch staatlicher Aufsichtsbehörden (im Rahmen der staatlichen Risikovorsorge) definiert. Sie baut auf den Kenntnissen aus klinischen Studien auf, in denen bei allen auftretenden gesundheitsrelevanten Problemen im Einzelfall medizinisch überprüft wird, ob die Einnahme des getesteten Arzneimittels eine Ursache ist (Kausalitätsbeurteilung von UAW-Verdachtsfällen). Ergebnissen solcher Einzelfallprüfungen wird eine hohe Validität attestiert. Da klinische Studien begrenzte Stichproben sind (eingeschlossene Population und Untersuchungszeitraum), können insbesondere UAW, die sehr selten, zeitlich verzögert oder nur bei spezifischen Risikogruppen (z. B. ältere Personen mit Komorbi-

ditäten und/oder Komedikationen) auftreten, nicht zuverlässig detektiert werden. Deshalb wird die Sicherheitsüberwachung auch nach der Zulassung fortgesetzt. Bisher stützt sich diese im Wesentlichen auf zwei Säulen:

- Durch sogenannte 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 Pharmaunternehmen produktbezogen und andererseits beim Bundesinstitut für Arzneimittelsicherheit (BfArM) produktübergreifend deutschlandweit geführt. Pharmaunternehmen sind verpflichtet, alle ihnen angezeigten UAW-Verdachtsfälle zeitnah an das zentrale Register beim BfArM zu melden. Parallel gibt es über die jeweiligen Berufsordnungen der Ärzte und Apotheker eine Selbstverpflichtung, UAW-Verdachtsfälle an die jeweilige Arzneimittelkommission zu melden, die diese Meldungen ebenfalls an das BfArM weiterleitet. Seit einigen Jahren können auch Bürger UAW-Verdachtsfälle dem BfArM direkt melden.

Mit beiden Instrumenten können 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 Arzneimiteleinsatz lediglich beobachtet. Weil in die normale medizinische Behandlung nicht eingegriffen

wird, fehlen oft spezifische medizinische Diagnosen zu individuellen biologischen Parametern, mit denen UAW-Verdachtsfälle überprüft werden könnten und sich die Kausalitätsbeurteilung fundieren ließe. Zur Hypothesen- bzw. UAW-Signalgenerierung sind sie zweifellos geeignet, jedoch wird je nach Sachlage im Einzelfall die Validität der Ergebnisse als nicht ausreichend bewertet. Teilweise werden weitere konfirmatorische Studien gefordert, die zusätzliche externe Belege für die Richtigkeit der Aussagen liefern sollen. (Die Revision der Nutzen-Risiko-Bewertung des Schmerzmittels Vioxx® weitete sich Anfang des Jahrtausends auch deshalb zum Skandal aus, weil Studienergebnisse immer wieder infrage gestellt und zusätzliche Belege durch neue Studien gefordert wurden [ausführlich z. B. in Ware 2005]).

- 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, Apotheker und/oder Patienten kaum erkannt. Die Medienaufmerksamkeit beeinflusst das Meldeverhalten erheblich: Wird über UAW-Vermutungen berichtet, steigt die Zahl der Meldungen deutlich (Problem sich selbsterfüllender Prophezeiungen). Bisher werden in Deutschland UAW-Verdachtsfälle überwiegend von Pharmafirmen gemeldet (ca. 85 % laut Zagermann-Muncke et al. 2010). Von Ärzten kommen nur 5 bis 10 % der Meldungen (AkdÄ 2005, S. 6). Durch unterschiedliche Meldekanäle des zentralen Registers beim BfArM sind Mehrfachnennungen nicht unwahrscheinlich. Oftmals gibt es nur lückenhafte Angaben zum Gesundheitszustand des Betroffenen, sodass eine Kausalitätsbeurteilung der UAW-Verdachtsfälle nur eingeschränkt möglich ist. Mögliche UAW-Risiken lassen sich aus Spontanmelderegistern nicht quantifizieren.

Mit den derzeitigen Instrumenten dauerte es laut Ware (2005) bisher im Mittel fünf 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.

Gesundheitsexperten betonen seit Jahren, dass eine systematischere Analyse medizinischer Versorgungsdaten die derzeitigen Instrumente der Pharmakovigilanzüberwachung ergänzen und verbessern können. Die US-amerikanische Food and Drug Administration (FDA) hat beim Vioxx-Skandal bereits vor mehr als zehn Jahren die Tür geöffnet für Analysen, die auf umfassenderen Datenbeständen aus der medizinischen Versorgung aufbauen und das UAW-Wissen ergänzen und fundieren können. 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.

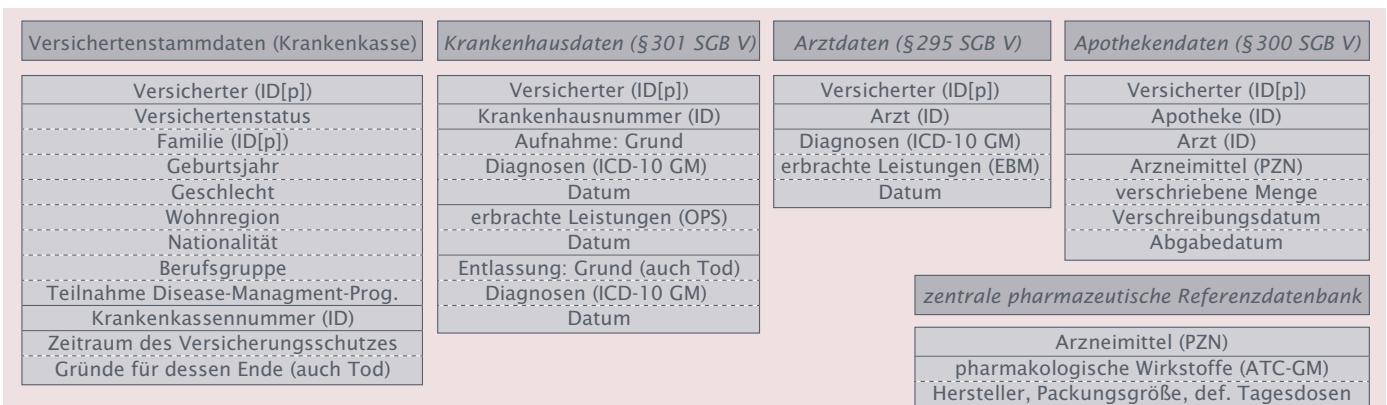
Medizinische Versorgungsdaten in Deutschland

In Deutschland werden Daten zur medizinischen Versorgung von Patienten in einer Vielzahl medizinischer Einrichtungen (ca. 2.000 Krankenhäuser, 150.000 Arztpraxen und 20.500 Apotheken) in Fall- bzw. Patientenakten zunehmend elektronisch erfasst und in Krankenhaus-, Praxis- bzw. Apothekeninformationssystemen geführt und verwaltet. Bisher sind diese Datenverwaltungssysteme einrichtungsspezifische Insellösungen. Diese personenbezogenen Gesundheitsdaten unterliegen in Deutschland grundsätzlich der ärztlichen Schweigepflicht (§ 230 StGB), die nur der Patient im Einzelfall aufheben kann, es sei denn, es gibt ein höherwertiges, gesetzlich legitimes Rechtsgut. Ein solcher Fall ist die Abrechnung medizinischer Versorgungsleistungen. Dafür werden aus den Patienten- bzw. Fallakten gesetzlich definierte einfallbezogene Abrechnungsdaten abgeleitet und an die jeweiligen Krankenkassen der Patienten (derzeit ca. 120 gesetzliche und ca. 40 private) übermittelt. Den normativen Rahmen für die Erhebung und Ver-

wendung dieser Abrechnungsdaten bilden grundsätzlich das Sozialgeheimnis, dem alle Krankenkassen unterliegen, sowie der Sozialdatenschutz. Auf dieser Grundlage werden die für die Abrechnung (Primärnutzung) notwendigen Datenstrukturen, Merkmale und Formate gesetzlich festgelegt. Auch wird der Rahmen für eine mögliche Weiterverwendung (Sekundärnutzung) festgelegt. Organisationsintern könnten sowohl medizinische Einrichtungen als auch Krankenkassen die Daten zu Forschungs- und Planungszwecken nutzen, sofern sie intern die dafür notwendigen Ressourcen aufbringen könnten. Zu diesen Zwecken kann auch eine Übertragung personenbezogener Daten an Dritte für bestimmte Vorhaben beantragt werden. Sie ist zeitlich zu begrenzen und wo zumutbar, muss eine Einwilligung der Betroffenen eingeholt werden.

Unter diesen normativen Vorgaben baut das Bremer Leibniz-Institut für Präventionsforschung und Epidemiologie (BIPS) in Kooperation mit vier gesetzlichen Krankenkassen und der Universität Bremen gegenwärtig die nationale pharmakoepidemiologische Forschungs-

Abb. 2 Datenmodell der nationalen pharmakoepidemiologischen Forschungsdatenbank (GePaRD)



ID[p]: Identifikationsschlüssel [pseudonymisiert]
 ICD-10 GM: International Classification of Diseases; Version 10 (deutsche Fassung)
 OPS: Operationen- und Prozedurenschlüssel
 EBM: einheitlicher Bewertungsmaßstab
 PZN: Pharmazentralnummer (Schlüssel zur zentralen pharmazeutischen Referenzdatenbank)
 ATC-GM: anatomisch-therapeutisch-chemische Klassifikation (deutsche Fassung)

Quelle: nach Schepers et al. 2015, S. 87

datenbank (»German Pharmacoepidemiological Research Database« [GePaRD]) auf, in die definierte Daten von ca. 17 Mio. gesetzlich versicherten Personen aus allen Bundesländern (ca. 20% der deutschen Bevölkerung) seit 2004 einfließen (Abb. 2). Statt die Einwilligung zur spezifischen sekundären Nutzung zu Forschungszwecken bei jedem einzelnen Versicherten einzuholen, wurde folgendes Verfahren vereinbart (Schepers et al. 2015, S. 88):

1. Die beteiligten Krankenkassen bereinigen und pseudonymisieren ihre codiert erhobenen und bereits qualitätsgeprüften Abrechnungsdatensätze anhand eines kasseninternen Schlüssels und übermitteln sie auf speziell geschützten Datenträgern jahrgangsweise an eine Vertrauensstelle an der Universität Bremen (vom BIPS organisatorisch getrennt).
2. Die Vertrauensstelle prüft die Datensätze mit Routinealgorithmen und klärt Unplausibilitäten mit den jeweiligen Krankenkassen. Anschließend werden die Datensätze ein zweites Mal pseudonymisiert und zu den Datenbeständen früherer Jahrgänge hinzugefügt. Gelieferte Daten, interne Pseudonymisierungsschlüssel und der Auswertungsdatenbestand werden getrennt aufbewahrt. Nur der Letztere wird an die Datenaufbereitungsstelle des BIPS weitergegeben. Dieses Verfahren führt zu gewissen zeitlichen Verzögerungen (die Daten von 2013 standen erst im Laufe des Jahres 2015 für wissenschaftliche Sekundäranalysen bereit).
3. Nur BIPS-Mitarbeiter dürfen eine konkrete Datennutzung bei den beteiligten Krankenkassen und zwei zuständigen Aufsichtsbehörden beantragen, die separat prüfen und gegebenenfalls genehmigen. Externe Wissenschaftler können sich über Kooperationen beteiligen.
4. Entsprechend der jeweiligen Genehmigung stellt die Datenaufbereitungsstelle des BIPS den Auswertungsdaten-

satz mit den definierten Dateninhalten dem verantwortlichen BIPS-Wissenschaftler bereit.

5. Dieser führt mit seinem Team die jeweiligen Analysen ohne Beteiligung der Vertrauens- und Aufbereitungsstelle durch.

Für die Suche nach möglichen UAW werden einerseits Arzneimittelaufnahmen als Ursache und andererseits gesundheitliche Probleme als mögliche Folge anhand codierter Abrechnungsdaten benötigt:

Aus den Abrechnungsdaten lassen sich keine Angaben zur tatsächlichen Arzneimittelaufnahme durch die Versicherten ableiten. Den besten Näherungswert bilden die Apothekenabrechnungsdaten (Arzneimittelabgabe an Versicherte). Sie enthalten nur Substanzen, die zum Leistungskatalog gesetzlicher Krankenversicherungen gehören (keine verschreibungsfreien Mittel). Die Apothekendaten müssen mithilfe pharmazeutischer Referenzdatenbanken teilweise umcodiert werden, um für derzeit ca. 6.500 Wirkstoffe bzw. Wirkstoffkombinationen den jeweiligen Einnahmezeitraum und die Dosierung abzuleiten.

Gesundheitliche Probleme können anhand von medizinischen Diagnosen definiert werden. Die dafür entwickelte »International Classification of Diseases« (ICD) klassifiziert gesundheitliche Probleme derzeit in ca. 12.000 Kategorien. Als mögliche unerwünschte Arzneimittelwirkung werden nur Probleme berücksichtigt, die während oder nach einer Arzneimittelaufnahme auftraten. Sowohl ambulante als auch stationäre medizinische Einrichtungen sind verpflichtet, die erbrachten medizinischen Leistungen mit der Angabe von Haupt- und Nebendiagnosen im jeweiligen Abrechnungsverfahren zu begründen. Inwiefern medizinische Leistungen legitimierende Diagnosen vollumfänglich und korrekt gesundheitliche Probleme von Versicherten abbilden, kann kaum zuverlässig abgeschätzt werden. Die ICD-Klassifika-

tion wird im Mehrjahresrhythmus den aktuellen Entwicklungen angepasst. Insbesondere für Längsschnittanalysen sind diese Überarbeitungen nicht unproblematisch, da sie zu Brüchen in den Datenbeständen im Zeitverlauf führen können.

Datenanalyse

UAW-Signale werden mit Disproportionalitätsanalysen von Kreuztabellen generiert, die für definierte Arzneimittel-Diagnose-Kombination aufgestellt werden (Abb. 3).

Abb. 3 Kreuztabelle

		Diagnose	
		ja	nein
Arzneimittel	ja	a	b
	nein	c	d

Je umfangreicher die Abrechnungsdatenbestände, desto genauer können die Werte für alle vier Felder ermittelt, daraus für jede Arzneimittel-Diagnose-Kombination unterschiedliche »Risikomaße« (z. B. Reported Odds Ratio [ROR] = (ad/bc)) berechnet und als Indikator für die Stärke eines UAW-Signals interpretiert werden. Bei sehr seltenen Ereignissen liefern klassische Risikomaße jedoch keine zuverlässigen Ergebnisse mehr (hohe Werte sind dann kein zuverlässiges UAW-Signal mehr). In diesem Fall können Bayes'sche Netze (basieren auf bedingten Wahrscheinlichkeitsverteilungen und sind vergleichsweise rechenintensiv) trainiert werden, die UAW-Signale zuverlässiger generieren. Die FDA setzt dieses Verfahren bei sehr seltenen Ereignissen ergänzend ein.

Mit den inzwischen verfügbaren Algorithmen können auch in umfangreichen Datenbeständen UAW-Signale sehr breit gesucht werden, d. h., es werden wenig bis

keine Restriktionen vorgegeben, 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 nach dem »anatomisch-therapeutisch-chemischen Klassifikationssystem« (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). Bei solch unbeschränkten Suchen in großen Datenbeständen ist damit zu rechnen, dass sehr viele UAW-Signale generiert werden, die im Anschluss gruppiert und bewertet werden in

- » bereits bekannte UAW-Signale (in nationalen und/oder internationalen Bulletins zu Arzneimittelsicherheit aufgeführt),
- » aus medizinischer Sicht sehr unplausible Zusammenhänge sowie
- » potenziell relevante UAW-Signale, die weiter beobachtet oder vertieft geprüft werden sollten.

Interpretation der Ergebnisse und Folgeaktivitäten

Laut Schepers et al. (2015, S. 92 f.) werden mit diesem Verfahren lediglich Hypothesen zu möglichen UAW generiert. Ziel sei es nicht, ein tatsächliches Risiko sicher festzustellen. Folglich ist bei allen UAW-Signalen zu prüfen, welche Maßnahmen der weiteren Risikoüberwachung resultieren sollten. Folgende Fälle werden unterschieden:

- » Ein potenzielles Sicherheitsrisiko deutet sich an, das UAW-Signal ist aber nicht so stark bzw. statistisch stabil, dass es bereits spezifische konfirmatorische Studien legitimieren könnte (z. B. bei sehr seltenen Ereignissen). Solche schwachen UAW-Signale können beispielsweise zunächst mithilfe von

sequenziellen Testverfahren niedrigschwellig weiter beobachtet werden. Wurden die UAW-Signale nur mit einem Teil der GePaRD-Daten erzeugt, könnte der andere Teil zur Signalprüfung mit weiteren Beobachtungsstudien genutzt werden (Schepers et al. 2015, S. 93).

- » Das UAW-Signal ist so deutlich, dass auf eine andere Art und Weise (z. B. auf der Basis anderer Datensätze) oder mit spezifischen konfirmatorischen Studien die Hypothese extern übergeprüft werden sollte.

Bisher gibt es keinen methodisch-fachlichen Konsens zur Bewertung der Ergebnishüte von UAW-Signalen und zur Ableitung von Folgemaßnahmen. Vielmehr wird im Einzelfall entschieden.

Situation in Deutschland und internationaler Vergleich

GePaRD-Datenbestände gelten normativ als nichtanonymisierte Sozialdaten. Bereits deren Übermittlung für Forschung und Planung »an Dritte« ist nur für vorab spezifizierte Vorhaben (Zweckbindung) mit festgelegtem Löschtermin (zeitliche Befristung) zulässig (§ 75 SGB X). Dadurch werden die Möglichkeiten, GePaRD für breite UAW-Suchläufe zu nutzen, beschränkt. Die Zweckbindung erfordert, dass zumindest Hinweise oder Vermutungen zu Arzneimittel-Ereignis-Zusammenhängen existieren müssen, die im Vorfeld zu spezifizieren sind (Schepers et al. 2015, S. 91). Aus dieser Perspektive können Hypothesen durch eine Datenanalyse untermauert, nicht aber generiert werden. So wurde beispielsweise durch eine Analyse des GePaRD 2014 die vorab bereits bestehende Hypothese der Verdopplung des Fieberkrampfrisikos bei einer Vierfachimpfkombination (Mumps, Masern, Röteln und Windpocken) im Vergleich zur klassischen Dreifachkombination und separater Windpockenimpfung erhärtet (Schink et al. 2014). Aufgestellt wurde diese Hypo-

these 2009 in den USA. Auch die Nutzungsbefristung beschränkt die derzeitigen Möglichkeiten. Dadurch können vor allem solche UAW, die mit erheblicher zeitlicher Verzögerung auftreten (z. B. karzinogene Folgen), kaum verfolgt oder nachgewiesen werden.

Auch in etlichen anderen Industrieländern werden seit einigen Jahren sogenannte elektronische Gesundheitsdaten, die aus der medizinischen Versorgung abgeleitet werden, für Forschungszwecke zugänglich gemacht (Schepers et al. 2015, S. 94). In den USA wurde 2007 die FDA-initiierte »Sentinel-Initiative« gesetzlich verankert, die sich zum Ziel setzt, elektronische Gesundheitsdaten von 100 Mio. Einwohnern für die aktive Überwachung von Arzneimitteln nach Marktzulassung zugänglich zu machen. Auch Großbritannien gehört zu den Vorreitern der Bereitstellung elektronischer Gesundheitsdaten für wissenschaftliche Zwecke. Bereits in den 1990er Jahren wurde mit dem Aufbau der britischen »General Practice Research Database« heute »Clinical Practice Research Datalink« begonnen. 2011 startete eine Regierungsinitiative, die durch die Intensivierung der forschungsseitigen Nutzung von elektronischen Gesundheitsdaten eine international herausgehobene Stellung im Gesundheitsforschungsbereich anstrebt. Die Datenbank stellt inzwischen medizinische Versorgungsdaten nahezu der gesamten englischen Bevölkerung (mehr als 50 Mio. Personen) für Forschungszwecke bereit. Weitere umfangreiche nationale Gesundheitsforschungsdatenbanken gibt es u. a. in den Niederlanden, in Dänemark, Estland, Italien und Spanien. Diese sechs europäischen Länder arbeiten inzwischen daran, ihre elektronischen Gesundheitsdatenbestände zu verlinken und gemeinsam zu erschließen.

Fazit

Die üblichen Prozessschritte jeglicher statistischer Analysen treffen auch auf Data-

Mining-Verfahren zu. Wie am Beispiel Pharmakovigilanz gezeigt wurde, wachsen gesundheitsbezogene Datenbestände bei unterschiedlichen Akteuren in erheblichem Umfang. In etlichen Industrieländern wird daran gearbeitet, diese gemeinsam zu erschließen. Der Weg zur Interoperabilität der Datenbestände im Gesundheitssystem ist jedoch weit und Normung, Codierung und Klassifikation flankieren ihn. Derzeit werden in etlichen Ländern Teilbestände in separate Forschungsdatenbanken kopiert und in unterschiedlicher Form zugänglich gemacht. In Deutschland wird der Zugang vergleichsweise restriktiv gehandhabt. Trotz doppelter Pseudonymisierung erhalten nur institutsinterne Wissenschaftler auf Antrag einen Datenzugang zur nationalen pharmakoepidemiologischen Forschungsdatenbank. Andere Länder sind bezüglich der wissenschaftlichen Nutzung von Gesundheitsdaten ambitionierter, insbesondere solche, die Open Knowledge (vom Zugang zu Daten über Ergebnisse bis zu wissenschaftlichen Publikationen) als Leitbild auch politisch erheblich forcieren. Dabei dürfte es unterschiedliche Ansichten geben, wann Daten als anonymisiert bzw. personenbezogen betrachtet werden.

Data-Mining-Analysen auch auf der Basis großer Datensätze gehen zumindest im Gesundheitsbereich auch heute noch mit genauer Datenerfassung, Aufbereitung und Validierung einher. Die Big-Data-These, dass sich mit sehr viel größeren Datenmengen das Bedürfnis nach Exaktheit überwinden ließe und unscharf gut genug wäre (Mayer-Schönberger/Cukier 2013, S. 20 f.), lässt sich zumindest anhand des Pharmakovigilanzbeispiels nicht untermauern. Wenn die Validität der Daten angezweifelt werden kann, werden die Ergebnisse infrage gestellt – umso mehr, je stärker dies Positionen und Anteile an den umsatzstarken globalen Pharmamärkten tangiert.

Auch wenn Analyseverfahren immer komplexer werden, zunehmend mehr Ar-

beitsschritte in Algorithmen internalisiert werden können und teilweise kaum noch nachvollziehbar erscheint, wie Ergebnisse zustande kommen, erfordert die Prüfung der Richtigkeit und die Bewertung der Validität nach wie vor erhebliches Fachwissen. Um dies zu erweitern und zu fundieren, muss jenseits der Beobachtung von Korrelationen auch zukünftig Kausalitätsfragen nachgegangen werden.

Katrin Gerlinger

Literatur

AkdÄ (Arzneimittelkommission der deutschen Ärzteschaft) (Hg.) (2005): Pharmakovigilanz. Sonderheft, Berlin, www.akdae.de/Arzneimitteltherapie/TE/A-Z/PDF/Pharmakovigilanz.pdf (17.5.2016)

Bernsdorf, B.; Bierbrauer, H.; Büscher, O.; Mütterthies, A.; Pakzad, K.; Wenzel, T.; Woditsch, S. (2015): Data-Mining: Gesellschaftspolitische und rechtliche Herausforderungen. Data-Mining mit Geodaten (Fallstudie 2). Münster

Bitkom (2015): Kognitive Maschinen – Meilenstein in der Wissensarbeit. www.bitkom.org/Publikationen/2015/Leitfaden/Kognitive-Maschinen/150213-Kognitive-Maschinen-11Febr2015.pdf (25.1.2016)

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. (1996): From data mining to knowledge discovery in databases. In: *AI Magazine* 17(3), S. 37–54

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

Mayer-Schönberger, V.; Cukier, K. (2013): *Big Data. Die Revolution, die unser Leben verändern wird.* München

Schepers, J.; Schlünder, I.; Drepper, J.; Semler, S.; Rüping, S.; Quix, C.; Stroet-

mann, K.; Rennoch, J. (2015): Data-Mining in der Medizin und im Gesundheitssystem – gesellschaftspolitische und rechtliche Herausforderungen. Gutachten, Berlin (unveröffentlicht)

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(6), S. 645–650

Ware, W. (2005): The Vioxx Saga: Perspective on the Recall. www.yourhealthbase.com/Vioxx.htm (22.3.2016)

Zagermann-Muncke, P.; Frölich, S.; Schulz, M. (2010): Unerwünschte Wirkungen an die AMK melden. In: *Pharmazeutische Zeitung* 10, <http://pharmazeutische-zeitung.de/index.php?id=32933> (22.3.2016)