Prof. Dr. med. Bertram Häussler

Von „Big Data“ zur Exploration von Assoziationen im Raum von Morbidität und Versorgung

Wie „Big Data“ in der Gesundheitsforschung diskutiert wird und welche Folgen dies hat

Zusammenfassung

Die Diskussion um „Big Data“ hat auch die deutsche Gesundheitswissenschaft erreicht und geprägt, obwohl sie aus der Beratung von Internetunternehmen stammt und dort zum Ziel hat, selbst generierte Daten verstärkt für kommerzielle Zwecke nutzen zu können. Die Rezeption der hiesigen Gesundheitsforschung hat sich jedoch auf den Aspekt konzentriert, ob es wissenschaftlich zulässig sei, Daten korrelativ miteinander in Beziehung zu setzen, zu deren Zusammenhängen man bisher keine Hypothesen hatte. In diesem Artikel wird dafür plädiert, trotz der vielfältigen Warnungen nicht die Möglichkeiten zu verpassen, die sich auch der deutschen Gesundheitsforschung durch die Nutzung großer Datensätze bieten, und eine breit angelegte Assoziationsforschung befürwortet.

Schlüsselwörter

Big Data, E-Commerce, Gesundheitswissenschaft, Assoziationsforschung

Crossref/doi

doi: 10.24945/MVF.03.18.1866-0533.2082

Von allen Diskussionen um „Big Data“ interessiert uns hier nur ein kleiner Ausschnitt: Die Vorstellungen, die zu diesem Thema in der deutschen Gesundheitsforschung expliziert worden sind oder die in impliziter Weise die Diskussion beeinflusst hatten. Diesen Ausschnitt eingehender zu untersuchen ist deshalb wichtig, weil die biomedizinische Forschung weltweit dabei ist, ihr Instrumentarium entscheidend zu erweitern. Der Begriff „Big Data“ leistet jedoch leider wenig, um diesen Prozess zu verstehen – im Gegenteil. Es geht um die historisch nie dagewesene Möglichkeit, sehr viele Menschen im Hinblick auf die natürliche Entwicklung ihrer Gesundheit zu beobachten, und dabei auch die Anwendung und Einwirkung massenhaft verfügbarer Interventionen zu studieren. Dieser Ansatz hat ein Potenzial, das noch nicht begriffen worden ist. Die Frage, welchen Zugang Deutschland zu dieser Entwicklung finden wird, stellt sich genau aus dieser Perspektive.

>> „Big Data“ ist kein Begriff, schon gar kein wissenschaftlicher. Die Kombination der beiden Wörter ist ein Buzzword, ein Schlagwort, ein Slogan. Er wurde in die Welt gesetzt – wie sollte es auch anders sein – von einer amerikanischen Unternehmensberatung. Gartner Inc. schickte 2011 folgenden Tweet in die Welt: “Big Data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.” Der Impuls dazu kam aus dem e-Commerce und ist in diesem Kontext zu verstehen: „In 2001/02, IT organizations must look beyond traditional direct brute-force physical approaches to data management. Through 2003/04, practices for resolving e-commerce accelerated data volume, velocity, and variety issues will become more formalized/diverse. Increasingly, these techniques involve tradeoffs and architectural solutions that involve/impact application portfolios and business strategy decisions.“ Wie können Firmen aus den Daten, die sich aus ihren e-Commerce-Aktivitäten entwickeln, mehr über ihre Kunden und ihre eigenen Prozesse lernen, daraus Nutzen für sich ziehen? Darum geht es bei „Big Data“[1]. Wir sollten den Begriff in der Gesundheitsforschung eigentlich nicht verwenden.

Aus diesem Kontext kommen die „3 V's“[2], die heute in vielen deutschen Publikationen zitiert werden. Mit drei Phänomenen sollen sich die Firmen befassen, wenn sie mehr aus ihren Daten machen wollen: Große Mengen müssen analysiert werden, Analysen werden schneller gebraucht als bisher üblich und die Beschränkung auf die schmale Kost der SQL-gebändigten Daten soll überwunden werden: Die Statistikprogramme konnten bis dato nur eine begrenzte Menge an Datensätzen verarbeiten – es gab aber zunehmend viel mehr. Die Analysen dauerten bisher eine Ewigkeit mit den verfügbaren Ressourcen – es musste aber viel schneller gehen (z. B. wenn ein konkreter Kunde im Internet auftauchte) und es gab zunehmend Informationen, die sich nicht der bekannten Architektur von SQL-basierten Programmen fügten – es wurde entdeckt, dass Bilder und Texte viel zu viel ungenutzte Informationen bargen als dass man sie ungenutzt liegen lassen sollte. Fotos zum Beispiel, die nunmehr auch in Bezug auf wichtige Eigenschaften automatisch klassifizierbar wurden (Beispiel siehe unten). „Volume“, „velocity“ und „variety“ muss man beherrschen, wenn man im e-Commerce mit Amazon mithalten wollte. „Big Data“ ist der Slogan, der geschaffen wurde, um den Beratungskunden das Problem klarzumachen: Große Fallzahlen mit verschiedenartigen Daten schnell zu analysieren – Hardware, Software und die Organisation des Wissens müssen daran angepasst werden. „Big Data“ ist bis in unsere Tage ein überaus erfolgreicher Slogan geworden. Daran ist nichts, was einen deutschen Gesundheitsforscher aufregen müsste.

Rezeption von „Big Data“ in der deutschen Gesundheitsforschung

Tut es aber trotzdem bis zum heutigen Tag. Dabei hängt die Aufregung eigentlich in der Luft, weil es so gut wie nichts Vergleichbares zum e-Commerce gibt im deutschen Gesundheitssystem. Es existieren keine gesundheitsbezogenen Plattformen, auf denen verschiedenste Datenquellen von Millionen von Menschen verfügbar wären. Es gibt keine Geschäftsmodelle, bei denen man Daten solchen Prozeduren unterziehen könnte, um schnellere und effizientere Entscheidungen zu treffen. Die Aufregung entzündet sich an der Fiktion, dass es diese Daten gäbe, und dass man sie für die Steuerung des Gesundheitssystems ähnlich nutzen könnte, wie Amazon sein Wissen nutzt, um die Kunden zu besseren Kunden zu machen. Wenn das so wäre hätte man es mit folgenden Herausforderungen zu tun:

— — — —

1: Svetlana Sicular hat darauf hingewiesen, dass es nicht die „3 V's“ sind, die die Definition von „Big Data“ ausmachen [15].
2: Manchmal ist auch schon von fünf V's die Rede

1. Missbrauch durch Manipulation und Diskriminierung von Plattformnutzern.
2. Ungerechte Verteilung von Einnahmen aus der Internetwirtschaft
3. Verschwendung von Ressourcen, weil die Analysen überwiegend Unsinn ergeben.
4. Ungerechte Monopolisierung von Informationen bei Internetfirmen.
5. Verstoß gegen verbreitete Regeln des wissenschaftlichen Denkens.

Die ersten drei Vorwürfe beziehen sich mehr oder weniger auf ethische Aspekte. Sie kritisieren ein Geschäftsmodell, das es im Gesundheitsbereich in Deutschland so nicht gibt (zumindest noch nicht). Auf diesen Komplex wird im Folgenden nicht weiter eingegangen, obwohl klar ist, dass die ethische Kritik an erfolgreichen Unternehmen der Internetwirtschaft im Kern eine politische ist, was zu einem erheblichen Maße zu der insgesamt ablehnenden Haltung gegen das Phänomen „Big Data" beiträgt. Der vierte Vorwurf drückt hingegen so etwas wie eine geheime Sehnsucht nach der Analyse der genannten Daten aus und ist damit auch ein Element der allgemeinen Ambivalenz diesem Thema gegenüber[3].

Unsere Aufmerksamkeit richtet sich vielmehr auf den zuletzt genannten Vorwurf, dass durch die Beschäftigung mit großen Datenmengen die wissenschaftliche Sorgfalt oder gar die Wissenschaft unter die Räder kommen könnten. Dieser Kritikpunkt ist von weitaus größerer Tragweite, weil – wie noch zu zeigen ist – damit auch eine Forschung betroffen ist, die heute bereits möglich wäre.

Worum es geht: Chris Anderson, ein Journalist aus den USA, hat 2008 in einem populären IT-Journal einen Artikel veröffentlicht, der unter dem Titel „The End of Theory: The Data Deluge Makes the Scientific Method Obsolete" einen vor allem in Deutschland viel zitierten Artikel. Das Stück greift plakativ die Position des wissenschaftlichen Mainstream an und postuliert respektlos, dass man in Zukunft nur noch aus den Daten lesen müsse, und sich nicht mehr damit abmühen müsse, ihnen mit Hilfe von Theorie ihre Geheimnisse zu entlocken. Anderson zitiert den damaligen Forschungschef von Google: „All models are wrong, and increasingly you can succeed without them" und kommt zu dem Schluss „Google was right".

Diese halbstarken Töne aus der Internetwirtschaft sind vor allem deshalb interessant, weil sie in der Szene der Gesundheitswissenschaftler in Deutschland Schaudern hervorrufen konnten: „Die unter der Überschrift Big Data angekündigten Möglichkeiten der neuen Welt missachten die üblichen Eckpfeiler von Wissenschaftlichkeit fast völlig. Es wird verkündet, dass das Zeitalter der Kausalität vorüber sei und nun nur noch Korrelationen wichtig seien. Auf Theoriebildung könne verzichtet werden, ausreichend Daten sind der Rohstoff und die Methode."[4] [2]

Mit der Proklamierung des „Petabyte Age", in dem die Gesetze der wissenschaftlichen Schwerkraft nicht mehr gelten sollen, ist Chris Anderson zum wissenschaftstheoretischen Luzifer aufgestiegen. Warum ein amerikanischer Fachjournalist deutsche Gesundheitswissenschaftler so ins Grübeln bringen kann, wird vielleicht dadurch verständlich, dass man großen Internetunternehmen doch eine ganze Menge zutraut, dass sie mit ihrer Annahme, nur mit Daten und ganz ohne Theorie die Welt verändern zu können, richtigliegen könnten. Und möglicherweise hat die Proklamation des „Correlation is enough" [1] ein wenig am herrschenden wissenschaftlichen Paradigma gerüttelt.

Ansonsten muss man konstatieren, dass „Big Data" zwar der Status einer selbständigen Methode zugewiesen bekommt, dass aber nicht präzisiert wird, was denn diese Methode eigentlich sein soll. Im Kern gehen die verschiedenen Definitionsversuche darauf zurück, dass Daten analysiert werden sollen, die man bisher nicht für geeignet ansah für gesundheitswissenschaftliche Analysen: Daten aus massenhaften internetbasierten Geschäftsprozessen. Diese Daten zu analysieren heißt aber nichts anderes, als nach Auffälligkeiten ihres gemeinsamen Auftretens zu suchen, nach Mustern, Verkettungen und dergleichen. Das ist tatsächlich der Kern von „Big Data". Alle V's beschreiben lediglich, dass dazu neue Analysewerkzeuge entwickelt werden, die mehr und unterschiedliche Daten in kürzerer Zeit verarbeiten können[5].

Der Kern von „Big Data": Transaktionsdaten

Ist an „Big Data" außer „big" also nichts Besonderes dran? Doch, es ist aber nicht die Geschwindigkeit oder die Vielfalt. Das Besondere kommt dadurch zustande, dass die Quelle der Information keine Befragungen, keine öffentlichen Statistiken, keine Zählungen oder andere bekannte Quellen sind, sondern dass sie direkt aus den verschiedenen Transaktionen[6] im Internet abgeleitet sind – geschäftlichen, sozialen, kulturellen. Diese mittlerweile zahllosen Quellen eint, dass sie Datenspuren hinterlassen, die Aufschluss geben können über die Motive von Nutzern im Allgemeinen und auch im Einzelfall und zwar ohne dass diese Daten für den Zweck der Fragestellung explizit erzeugt worden sind. Ein Universum von Fragestellungen mit unendlich vielen sozial-, wirtschafts- und auch gesundheitswissenschaftlichen schließt sich daran an.

Das Buch „Dataclysm" hat daher auch den treffenden englischen Untertitel „Who We Are When We Think No One's Looking." [13] Es zeigt die Vielfalt und den Bedeutungsgehalt solcher Analysen am Beispiel der Daten einer internetbasierten Partnervermittlung[7]. Zwei Beispiele zeigen dies:

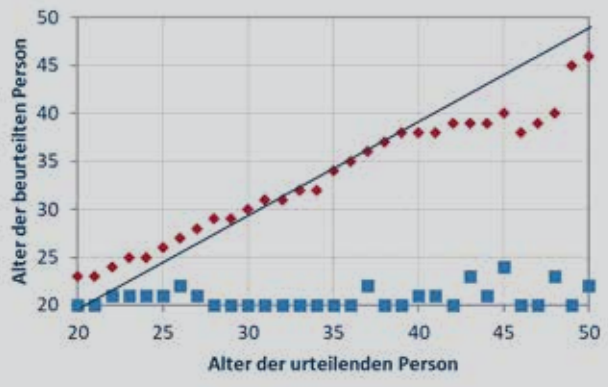
1. Bis zum Alter von 40 finden Frauen die Männer am attraktivsten, die ungefähr in ihrem eigenen Alter sind. Nach dem 40. Lebensjahr liegt das attraktivste Alter leicht darunter. Dagegen finden Männer zwischen 20 und 50 Jahren weitgehend uniform 20-jährige Frauen am attraktivsten.
2. Weiße Frauen wie auch weiße Männer finden potenzielle Partner schwarzer Hautfarbe deutlich unterdurchschnittlich attraktiv.

Die Analysen legen Verhaltensweisen und Einstellungen offen, zu denen sich Befragte normalerweise niemals bekennen würden, weil es politisch nicht korrekt ist, solche Einstellungen zu haben. Welcher Mann würde zugeben, dass er insgeheim nur auf 20-jäh-

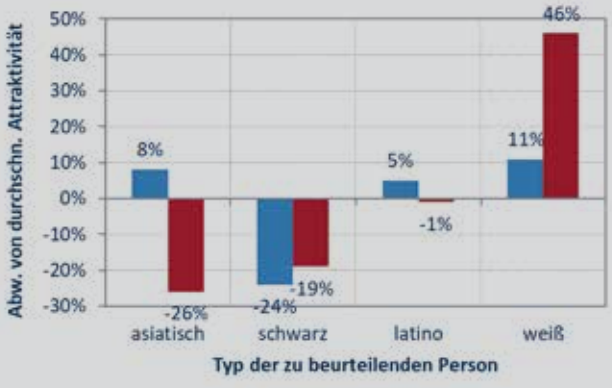
— — — —

3: Diese Ambivalenz kommt auch darin zum Ausdruck, dass die Klage über die Irrelevanz der Ergebnisse (Punkt 3) offensichtlich nicht mit anderen Punkten kompatibel sind wie z.B. Punkt 1.
4: Diesen Vorwurf macht Antes fälschlicherweise auch Viktor Mayer-Schönberger and Kenneth Cukier
5: Die Artikel, aus denen die „V's" zitiert sind, beschreiben aber keine Methode, sondern die wirtschaftliche Herausforderung, vor die sich die Anwender gestellt sehen.
6: Eine Transaktion im Sinne dieser Abhandlung ist die Übergabe von Information im Rahmen von definierten digitalen Prozessen. Dies umfasst alles von der willentlichen bis hin zur nicht willentlichen Aktivität, unabhängig vom Format und dem Umfang der daraus erzeugten Datenpakete.
7: OkCupid. Die Plattform hatte nach eigenen Angaben 2014 ca. 11 Millionen Nutzer.

Analysen zur Attraktivität von potenziellen Partnern



Welches Alter Männer bzw. Frauen beim anderen Geschlecht attraktiv finden in Abhängigkeit vom eigenen Alter (rot: attraktivstes Alter aus Sicht der Frauen; blau: attraktivstes Alter aus Sicht der Männer)



Einfluss der Typs[9] des jeweiligen anderen Geschlechts auf die Abweichung von der durchschnittlichen Attraktivität. (rot: Urteile weißer Frauen; blau: Urteile weißer Männer)

Abb. 1: Analysen zur Attraktivität von potenziellen Partnern. Quelle: Eigene Analysen nach [13].

rigen Frauen steht? Welche Frau würde zugeben, dass sie schwarze Männer weit weniger attraktiv findet als weiße?[8] Wer würde bestreiten, dass solche Erkenntnisse brauchbar sind für die Erforschung rassistischer Einstellungen und dass sie eine Lücke unseres Wissens füllen können?

Die zitierten Beispiele sind ein winziger Ausschnitt aus den Möglichkeiten für Analysen von Transaktionsdaten aus dem Internet. Alle großen Internetunternehmen bieten mittlerweile solche Daten auch zur wissenschaftlichen Nutzung an oder haben eigene Einheiten[10]. „Like Google, a gaggle of social media networks such as Facebook, Twitter, LinkedIn, Foursquare, and others sit on an enormous treasure chest of datafied information that, once analyzed, will shed light on social dynamics at all levels, from the individual to society at large." [10] Dennoch muss davon ausgegangen werden, dass nur ein Bruchteil der Information das Licht der Öffentlichkeit erblickt. Der größte Teil dürfte im Rahmen privatwirtschaftlicher operativer Zwecke verwendet werden.

Was die Beispiele aber auch zeigen ist der Umstand, dass die Analysen nicht auf den klassischen Wegen von protokollierten sozialwissenschaftlichen Untersuchungen zustande gekommen sind. Dennoch sollte man sich nicht der Vorstellung hingeben, hier werde alles mit allem korreliert und jeder sinnlose Zusammenhang kolportiert[11]. Natürlich kennen auch die Analysten aus der Internetwirtschaft das wissenschaftlich verheerende Potenzial des Confounding und entwickeln sophistizierte Methoden zur Eindämmung desselben. [13]

Warum „Correlation is enough" die Gemüter so erhitzt

Die Beispiele zeigen auch, dass die Suche nach Zusammenhängen zwischen Phänomenen den Kern der Verwendung von Transaktionsdaten aus dem Internet – also von „Big Data" – darstellt. „Correlation is enough" ist die übertreibende Beschreibung dessen. „Korrelationsbasierte Exploration" wäre die besser zutreffende Bezeichnung für dieses Tun. Oder: „Die Suche nach Auffälligkeiten." An dieser Stelle kommt der Ansatz allerdings mit dem wissenschaftlichen Grundverständnis der Gesundheitswissenschaft in Konflikt.

Das herrschende Paradigma in der Gesundheitsforschung ist ausführlich in den zahlreichen Memoranden zur Versorgungsforschung nachzulesen. Diese sind durch das „Netzwerk Versorgungsforschung" initiiert worden, das praktisch die gesamte akademische Gesundheitswissenschaft organisiert. Obwohl man der Tiefe der in den Memoranden vertretenen wissenschaftlichen Positionen mit ein paar Sätzen nicht gerecht werden kann, kann doch herausgelesen werden, dass sich „die Gesundheitswissenschaft" auf die erkenntnistheoretische Position des „kritischen Rationalismus" einigen kann. Dieser hatte sich bekanntlich in der Auseinandersetzung mit dem Positivismus gebildet und sich in dieser sehr alten Debatte auf die Seite der Deduktionisten geschlagen. Im Unterschied zum Induktivismus gilt dem kritischen Rationalismus das mit den Sinnen erfahrbare Datum praktisch nichts, weil es nicht gedacht werden kann, ohne dass man Begriffe dafür einsetzt, die bereits theoretische Annahmen über das Funktionieren der Welt enthalten. [12] Im „Memorandum IV" liest es sich so: „Theorie ist die Voraussetzung für den wissenschaftlichen Blick der empirischen Forschung. Empirische Daten erhalten ihre Relevanz nur über Theorie." [4]

Die vorgefundene erkenntnistheoretische Konstitution der Gesundheitswissenschaften hat ihre stärksten Auswirkungen auf die Frage, wie mit Daten umgegangen werden soll. Darf man sie einfach so nehmen wie sie sind und analysieren, ohne dass man vorab die exakte Frage stellt, wonach man denn eigentlich sucht und wann man davon reden darf, dass man die gestellte Frage beantwortet hat? „Nein", sagen die Gesundheits- und Versorgungsforscher: „In allen Studien zur Versorgungsforschung müssen Zielvariablen, Endpunkte bzw. Entscheidungskriterien weitestgehend a priori defi-

— — — —

8: Asiatische Männer rangieren noch weiter hinten.
9: Im Original ist von „race" die Rede.
10: Z. B. Facebook Insights, Twitter APIs, .
11: W. Bartens in der „Süddeutschen Zeitung" vom 18.3.2018: „So entwickelte sich der Käsekonsum pro Kopf von 2000 bis 2010 parallel zur Anzahl der Menschen, die auf tragische Weise unter ihrer Bettdecke erstickt sind. Eine Zehn-Jahres-Korrelation besteht auch zwischen den von 1999 bis 2009 in einem Pool Ertrunkenen und der Anzahl jährlicher Filmrollen von Nicolas Cage."[3]

niert, begründet, nachvollziehbar dokumentiert sowie transparent dargestellt werden." [11][12]

Es leuchtet ein, dass in diesem wissenschaftlichen Weltbild der Korrelation als Kern der Suche nach dem Zusammenhang allenfalls eine nachrangige Bedeutung zugemessen wird. In der Kritik an „Big Data" wird insinuiert, dass es sich bei Analysen, die auf Korrelationen basieren, nicht um seriöse Analysen handele, dass der wissenschaftliche Anspruch auf Kausalität durch die Korrelation herausgefordert würde.[2]

Diese Sorge findet sich auch in Bezug auf die Versorgungsforschung: „... die Big-Data-Diskussion rührt an ihrem historischen Begründungszusammenhang; anderenfalls verkümmert die Versorgungsforschung zur bloßen Datenakkumulation." [14] Es passt zu dieser distanzierten Position, dass sich die Versorgungsforschung überwiegend warnend, jedoch nicht offensiv um die Transaktionsdatenforschung gekümmert hat. Dies wird auch im Memorandum IV deutlich: „Angesichts der informationstechnologischen Revolution, die die Medizin nicht nur im Hinblick auf die Explosion der Datenmengen erfassen wird, stellt das Thema ‚Big Data' auch die Wissenschaftskonzeption und die Theoriebildung der Versorgungsforschung vor völlig neue Herausforderungen, die wir noch nicht durchdrungen haben (...)." [4]

Die Distanz hat letztlich zur Folge, dass in Gesundheits- und Versorgungsforschung solche Ansätze zu kurz kommen, die sich dem Gesundheitssystem explorativ nähern, und die sich – im Unterschied zu experimentellen Ansätzen – nicht der Testung von Hypothesen verschreiben, sondern der Generierung von neuen Erkenntnissen. Auch in dieser Hinsicht ist den Autoren des geraden zitierten Memorandums zuzustimmen: „Theorie benötigt Phantasie, und diese Phantasie bedient sich oftmals völlig unkalkulierbarer Quellen und Anregungen, womöglich am allerwenigsten der starren Gerüste einer formalisierten Hypothesentestung." [4]

Assoziationsforschung – ein unbegangener Weg zu neuen Erkenntnissen

Die Exploration von Internet-Transaktionsdaten könnte im Gesundheitsbereich an vielen Stellen hilfreich sein, wenn sie im deutschen Versorgungskontext verfügbar wären[13]. Daher stellt sich die Frage, welche Möglichkeiten Transaktionsdaten liefern, die zwar nicht aus Internetanwendungen stammen sondern aus Geschäftsprozessen der gesetzlichen Krankenversicherung in Deutschland. Die als „Routinedaten" bekannten Daten liegen ebenfalls massenhaft vor, und haben ihre Stärke im Bereich der Morbidität und der Interventionen, weniger im Bereich der Präferenzen und Wahlentscheidungen der Versicherten bzw. Patienten. Mit ihnen ist durchaus eine lebhafte Forschung entstanden, die das Bild von der Versorgung um einige Aspekte bereichert hat[14].

Weitgehend unbearbeitet ist jedoch der Aspekt geblieben, dass die Versorgung einer großen Population im Grunde ein riesiges natürliches Experiment darstellt: Millionen Menschen leiden unter Erkrankungen, die mit tausenden von Diagnosen kodiert werden, und werden in zahllosen Episoden ebenfalls tausenden verschiedener Therapien ausgesetzt. Ein großer Teil der Behandlungen dürfte keine Rätsel aufgeben, weil genau das abgebildet wird, was die Leitlinien oder das Lehrbuch fordern.

Allerdings dürften sich auch deutliche Varianzen zeigen, die sich in der wirklichen Welt deshalb ergeben, weil nicht alle Menschen gleich reagieren, weil sie genetisch unterschiedlichen Phänotypen angehören. Diese Varianzen werden sich in den Daten abbilden, auch wenn wir die Gründe dafür aufgrund mangelnder Daten nicht ohne weiteres erkennen können. Wir wissen heute, dass solche Effekte einen erheblichen Einfluss auf das Versorgungsgeschehen und letztlich auch auf den Erfolg von Therapien haben können. Darüber sollte man viel mehr wissen.

In einer großen Population wird man finden, dass viele Patienten Mustern von Komorbiditäten zuzuordnen sind, die auf der Basis des medizinischen Wissens zu erwarten sind. Das metabolische Syndrom ist hierfür ein geläufiges Beispiel. Die Entschlüsselung der genetischen Steuerung von zellulären Prozessen und der molekularen Grundlagen der Steuerung der Zellfunktionen lehrt uns zunehmend, dass einzelne genetische Konstellationen und Signalprozesse bei teilweise sehr unterschiedlichen Erkrankungen involviert sind, die man bisher nicht im Zusammenhang gesehen hat. In großen Populationen ergibt sich aufgrund der großen statistischen Power die Chance, solche seltenen Zusammenhänge entdecken zu können. Daraus könnte man ein wertvolles Instrument auch für die Erforschung biomedizinischer Zusammenhänge entwickeln. Transaktionsdaten aus der GKV könnten helfen, biomedizinische Hypothesen zu stützen oder zu verwerfen.

Darüber hinaus können wir davon ausgehen, dass applizierte Therapien nicht nur ihre intendierten Wirkungen entfalten, also auf die Erkrankung einwirken, für die sie zugelassen sind. Ihre Wirkung wird zum einen durch Komorbiditäten beeinflusst sein, wozu teilweise Wissen aus klinischen Tests vorliegt, teilweise aber auch nicht, weil die Zahl der dort getesteten Menschen zu gering ist. Zum anderen jedoch dürften die applizierten Therapien auch Wirkungen entfalten, wo es bisher niemand erwartet hat. Damit ist nicht der „off label use" in wenigen Einzelfällen gemeint, sondern die breite Wirkung bei anderen Krankheitszuständen. Es könnte sich daraus z.B. die Chance ergeben, lange bekannte Arzneimittel für die Behandlung von Krankheitszuständen einzusetzen, für die sie ursprünglich nicht gedacht waren. Auch dazu sollten wir mehr wissen, weil das „repurposing" von Arzneimitteln unter Umständen Fortschritte bei der Suche nach Therapien helfen könnte, wo wir heute noch im Dunkeln tappen (z.B. Demenz [8] [6]).

Diese skizzierten Anwendungen könnten Teil einer breit angelegten Assoziationsforschung sein, die – wenn man wollte – ohne weitere als „Big Data"-Anwendung im oben genannten Sinne bezeichnen könnte. Charakteristika wären die Herkunft der Daten aus Transaktionen bzw. Routineprozessen und ihre sehr große Zahl. Prägend ist auch der Umstand, dass man es hier ganz explizit nicht mit experimenteller Forschung zu tun hat, sondern mit klassischen Beobachtungsstudien, bei denen die Korrelation und die Beherrschung des Confounding mit statistischen Mitteln die methodische Wahl sind. Aber warum sollte es nicht möglich sein, das Potenzial der Amazon-Analytik auf die Gesundheitsforschung zu übertragen?

Könnte der Algorithmus für „Kunden, die diesen Artikel gekauft

— — — —

12: „Ausnahmen", z.B. zur Generierung von Hypothesen, seien jedoch erlaubt.

13: Hierzu haben Wegscheider und Koch-Gromus zahlreiche Anregungen gegeben. [16].

14: Auch sind erste Schritte unternommen, diese Daten einem breiteren Kreis zur Verfügung zu stellen [5].

haben, kauften auch …“ nicht auch bei der Aufgabe „Patienten, die Krankheit X haben, haben auch Krankheit …“ gute Dienste leisten?

An anderen Orten sind solche Projekte schon gestartet worden, auch unter Verwendung einer Kombination aus klinischen und genetischen Daten. [7]

Die Begründung für ein solches Programm kommt aus dem Ziel, neues Wissen zu erschließen, was immer der Antrieb für wissenschaftliche Veränderungen war. Paradigmenwechsel haben nie stattgefunden, weil durch ein Experiment eine Hypothese falsifiziert werden konnte, sondern weil ein neuer Ansatz viele neue Forschungsprojekte und viele neue Erkenntnisse versprach. [9]

Am Wissenschaftsverständnis des kritischen Rationalismus können im Übrigen all jene getrost festhalten, die in der klinischen Forschung daran arbeiten, den Irrtum aus den Ergebnissen zu verbannen. Aber es gibt viel mehr Pfade zu neuem Wissen, die wir uns durch methodische Orthodoxie nicht blockieren sollten. „Enough“ haben wir noch lange nicht. <<

From „Big Data“ to the Exploration of Associations in the Field of Morbidity and Health Care

The discussion on „big data“ has reached the discourse of German health science and has left its impact here, even though the term itself originates from consulting internet companies and describes a way to increase value from internet transactions. The national reception of the term focuses solely on the aspect, whether or not it would be scientifically acceptable to correlate data with each other, without having had established prior hypotheses for their relationship. This article is well aware of the numerous methodological concerns. Despite these warnings, it claims that one should take the opportunities of using large data sets for a broadly based association research in health and healthcare research.

Keywords

Big data, e-commerce, health science, healthcare research, association research.

Autorenerklärung

Der Autor erklärt, dass keine Interessenkonflikte vorliegen.

Zitationshinweis

Häussler, B., „Von Big Data zur Exploration von Assoziationen im Raum von Morbidität und Versorgung“ in: „Monitor Versorgungsforschung“ 03/18, S. 54-58, doi: 10.24945/MVF.03.18.1866-0533.2082

Literatur

1. Anderson C (2008) The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. In: Wired (https://www.wired.com/2008/06/pb-theory/), [Abruf am: 12. Mai 2018]
2. Antes G (2016) Ist das Zeitalter der Kausalität vorbei? In: Z. Evid.Fortbild.Qual.Gesundh.wesen (ZEFQ) S16—S22
3. Bartens W (2018) Dr. Data. Süddeutsche Zeitung 18. März 2018. http://www.sueddeutsche.de/wissen/medizin-dr-data-1.3908072 [Abruf am: 12. Mai 2018].
4. Baumann W, Farin E et al. (2016) Memorandum IV: Theoretische und normative Fundierung der Versorgungsforschung. Gesundheitswesen 78(05), 337-352.
5. Bundesministerium für Wirtschaft und Energie (2017): D066-01 SmartData – SAHRA. SAHRA: Rechtssichere Plattform für Versorgungsdaten aus verschiedenen Behandlungskontexten. 15. Februar 2017. http://www.tmf-ev.de/Themen/Projekte/D06601SmartDataSAHRA.aspx [Abruf am: 12. Mai 2018].
6. Decourt B, Lahiri DK et al. (2017) Targeting Tumor Necrosis Factor Alpha for Alzheimer's Disease. Current Alzheimer research 14(4), 412-425
7. Denny JC, Bastarache L et al. (2013) Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. In: Nat Biotechnol. 2013 December ; 31(12): 1102–1110
8. Ekert JO, Gould RL et al. (2018) TNF alpha inhibitors in Alzheimer's disease: A systematic review. International journal of geriatric psychiatry 33(5), 688-694.
9. Kuhn TS. (1996). Die Struktur wissenschaftlicher Revolutionen. Frankfurt am Main. Suhrkamp
10. Mayer-Schönberger V, Cukier K (2013) Big Data: A Revolution That Will Transform How We Live, Work, and Think. New York Houghton Mifflin Harcourt.
11. Pfaff H, Glaeske G, et al. (2009) Memorandum III: Methoden für die Versorgungsforschung (Teil I) In: Das Gesundheitswesen 71(8/09), 505-510
12. Popper KR (1972) Die Logik der Sozialwissenschaften. In: Adorno TW, Dahrendorf R et al. (1972) Der Positivismusstreit in der deutschen Soziologie. Luchterhand
13. Rudder C (2014) Dataclysm. Who We Are When We Think No One's Looking. Crown Publishers
14. Schrappe M (2016) Versorgungsforschung braucht eine Digitale Agenda. Monitor Versorgungsforschung 9(2), 52-57
15. Sicular S (2013) Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three „V“s. https://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/#4fbe153b42f6 [Abruf am: 12. Mai 2018]
16. Wegscheider K, Koch-Gromus U (2015) Die Versorgungsforschung als möglicher Profiteur von Big Data. Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz 58(8), 806-812

Prof. Dr. med. Bertram Häussler
ist Leiter der IGES Gruppe, einem Verbund von Forschungs- und Beratungsunter-nehmen aus dem Bereich Infrastruktur und Gesundheit. Er ist Mediziner und Soziologe und lehrt als Honorarprofessor an der TU Berlin das Spezialgebiet „Ökonomik der pharmazeutischen Industrie“. Seine wissenschaftlichen Schwerpunkte sind Industrieökonomik, Versorgungsforschung, Gesundheitsökonomie, Nutzenbewertung, Digitalisierung im Gesundheitswesen und Arzneimittelentwicklung. Kontakt: BH@iges.com

