Bundesgesundheitsbl 2015 · 58:785–787
DOI 10.1007/s00103-015-2194-6
Online publiziert: 26. Juni 2015


**Iris Pigeot[1] · Svenja Jacobs[1] · Uwe Koch-Gromus[2]**
[1] Leibniz-Institut für Präventionsforschung und Epidemiologie – BIPS, Bremen, Deutschland
[2] Universitätsklinikum Hamburg-Eppendorf, Hamburg, Deutschland

# Große Datensammlungen im Gesundheitswesen – Chance oder Risiko?

Als in den USA im Jahr 2009 die H1N1-Epidemie, die sogenannte Schweinegrippe, ausbrach, hatte Google kurz zuvor ein wichtiges Ziel erreicht. Es war ihnen gelungen, auf Basis ihrer riesigen Datenbestände ein geeignetes mathematisches Modell zu ermitteln, das Google-Suchbegriffe identifiziert, deren häufige Nutzung den Verlauf von Grippeepidemien vorhersagen kann. Anhand der Verwendungsmuster der identifizierten Suchbegriffe war es den amerikanischen Gesundheitsbehörden möglich, die räumliche Ausbreitung der H1N1-Epidemie in Echtzeit abzubilden und geeignete Vorkehrungsmaßnahmen zu treffen.

Auch wenn eine in Science veröffentlichte Studie der Northeastern University Schwächen in der allgemeinen Vorhersagequalität des Algorithmus aufzeigt, ist dies trotz allem ein eindrucksvolles Beispiel für Big Data im Gesundheitswesen, das Viktor Mayer-Schönberger und Kenneth Cukier in ihrem Bestseller „Big Data“ beschreiben. Big Data wird in der Fachliteratur häufig mit den sogenannten drei „V“ in Verbindung gesetzt: große Datenmengen (Volume), die kaum oder sehr heterogen strukturie rt vorliegen (Variety) und die mithilfe mathematisch-statistischer Verfahren und Optimierungsalgorithmen möglichst flexibel und in Echtzeit verarbeitet werden (Velocity). Big Data bezieht sich damit nicht nur auf große Datensammlungen, sondern ebenso auf die angewendeten Auswertungsmethoden, die grundsätzlich nicht hypothesengetrieben sind. Big Data zielt darauf ab, oftmals durch die Anwendung von Data-Mining-Methoden, in möglichst kurzer Zeit Korrelationen, also Zusammenhänge, in Datenbeständen zu identifizieren. Sehr häufig wird dabei gleichzeitig auf unterschiedliche, nicht grundsätzlich qualitätsgeprüfte Quellsysteme zurückgegriffen. Damit unterscheidet sich Big Data von groß angelegten Forschungsdatenbanken, die Daten, in der Regel strukturiert und qualitätsgeprüft, als Basis hypothesengetriebener Forschung bereitstellen. Zwischen diesen beiden Extremen liegen große Datenbanken, die nicht primär zu Forschungszwecken angelegt werden, wie etwa Abrechnungsdaten der Krankenversicherungen, die aber dennoch ein großes Potenzial für die Forschung bieten.

In Deutschland ist die Nutzung von Big Data durch die Wissenschaft rechtlich deutlich stärker eingeschränkt als beispielsweise in den USA. So ist es unter anderem nicht möglich, auf Basis von Krankenversicherungsdaten mithilfe von Data-Mining-Methoden Arzneimittelrisiken im alltäglichen Gebrauch aufzudecken. Dennoch ist Big Data auch aus dem deutschen Gesundheitswesen kaum noch wegzudenken. Zahlreiche Apps und Geräte wie Fitnessarmbänder sammeln Biosignale oder andere gesundheitsrelevante Daten etwa zum Ess- oder Bewegungsverhalten. Sie geben den Nutzerinnen und Nutzern kurzfristig Rückmeldung zu ihrem Verhalten oder ihrem Gesundheitszustand, und das häufig nicht nur auf Basis deren eigener Daten, sondern ebenso auf Basis der Daten weiterer Personen oder anderer Quellen.

In der hypothesengestützten Gesundheitsforschung spielen große Datensammlungen in nahezu allen Bereichen eine große Rolle. So nimmt etwa die Genetik eine immer wichtigere Stellung ein: Über 20.000 Gene werden regelmäßig untersucht, inwieweit sie Krankheiten und Gesundheit beeinflussen. Ein anderes Beispiel ist die elektronische Gesundheitskarte, über die zukünftig die Behandlungsdaten aller gesetzlich Versicherten, die einer Speicherung nicht widersprechen, auf einer Serverplattform erfasst werden sollen. Ziel ist es, durch die zusätzlichen Informationen, die dem Arzt oder der Ärztin zur Verfügung gestellt werden, die Gesundheitsversorgung zu verbessern. Auch im Rahmen epidemiologischer Studien werden große Datenbanken aufgebaut, in denen nicht nur medizinische Parameter, sondern auch Lebensstilfaktoren und soziodemografische Variablen von Personen erfasst werden, um so Ursachen von Erkrankungen besser erforschen zu können. Aktuell startet mit der Nationalen Kohorte die bisher größte Gesundheitsstudie in Deutschland. Insgesamt sollen von 200.000 Personen, die ihr Einverständnis gegeben haben, Informationen zu Lebensstil, Vorerkrankungen und Risikofaktoren häufiger Volkskrankheiten wie Herz-Kreislauf-Erkrankungen, Diabetes, Krebs oder Demenz gesammelt und anonymisiert ausgewertet werden.

Für die Wissenschaft besitzen große Datensammlungen ein enormes Potenzial zur Beantwortung komplexer Fragestellungen. Die langfristigen Auswirkungen von Ereignissen wie Lärmbelastungen oder eines geringen sozioökonomischen Status in unterschiedlichen Altersstadien auf die Gesundheit im gesamten Lebensverlauf oder die Zusammenhänge zwischen Lebensstil, Genetik und dem Auftreten von Krankheiten wie Demenz oder Krebs sind hier nur zwei von vielen Beispielen. Aus den gewonnenen Ergebnissen lassen sich dann Handlungsempfehlungen und Leitlinien ableiten, die einen

direkten Nutzen für die Bevölkerung und das sie behandelnde medizinische Fachpersonal haben. Gleichzeitig ebnen große Datensammlungen den Weg zu einer personalisierten und selbstbestimmteren medizinischen Versorgung. Mithilfe moderner Medien sind die Menschen in der Lage, ihren Gesundheitszustand eigenständig zu kontrollieren. Dabei könnten z. B. individuelle Grenzwerte für Blutdruck oder Blutzucker festgelegt werden.

Alle genannten Beispiele zeigen, dass einerseits große Datensammlungen aus dem Gesundheitswesen gar nicht mehr wegzudenken sind. Andererseits steht ein großer Teil der Bevölkerung der zunehmenden Datensammlung in allen Lebensbereichen durch eine Reihe aktuell aufgetretener Skandale im Umgang mit sensiblen Daten, wie etwa dem NSA-Überwachungsskandal, teils unsicher, teils kritisch gegenüber. Umso mehr stehen alle Akteure in der Pflicht, einen sicheren und transparenten Umgang mit großen Datensammlungen zu garantieren. Dazu gehören, wenn möglich, eine weitgehende Anonymisierung der Daten und die Anwendung von Speicher- und Verschlüsselungslösungen nach höchsten technischen Standards, aber auch die Festlegung von Rahmenbedingungen, unter denen verschiedene Datenquellen miteinander verknüpft werden dürfen, und transparente Regeln, wer die gewonnenen Erkenntnisse zu welchem Zweck nutzen darf. Auch die Frage, wie dem Konflikt zwischen sinnvoller Datensammlung und unrechtmäßiger Vorratsdatenspeicherung Rechnung getragen werden soll, gilt es zu beantworten.

Das zu Beginn aufgeführte Beispiel der H1N1-Epidemie zeigt, dass auch die Gesundheitspolitik und die entsprechenden Behörden von der Nutzung von Big Data profitieren können. Ihre Rolle ist in Deutschland aktuell aber insbesondere darin zu sehen, Diskussionen zu unterstützen und die Nutzung großer Datensammlungen durch eine Gesetzgebung zu untermauern, die gleichermaßen relevanten datenschutzrechtlichen Bedingungen Rechnung trägt wie auch eine Forschung im Interesse und zum Nutzen der Bevölkerung möglich macht, statt diese zu verhindern.

Das hier vorgelegte Themenheft setzt sich in 11 Beiträgen mit den Chancen und Herausforderungen großer Datensammlungen für die gesundheitsbezogene Forschung auseinander. Dem Thema Big Data kommt dabei eine zentrale Rolle zu, die Behandlung des Themas geht aber darüber hinaus.

Die ersten vier Beiträge behandeln das Thema Big Data aus einer übergreifenden Perspektive. In einem programmatischen Beitrag vertritt *V. Mayer-Schönberger* die Position, dass Big Data-Ansätze einen bisher kaum vorstellbaren beschleunigten Zugang zu wissenschaftlichen Erkenntnissen, insbesondere in den Bereichen der Lebens- und Sozialwissenschaften, ermöglichen werden. Dabei werde es über die Weiterentwicklung von digitalen Techniken bei der Sammlung und Analyse von Daten auch zur Veränderung wissenschaftlicher Erkenntnismethoden insgesamt kommen. Der Autor warnt aber auch vor den Gefahren einer missbräuchlichen Verwendung von Big-Data-Ansätzen.

Auch *S. Rüping* sieht das Gesundheitswesen einschließlich der Forschung in der Medizin als einen Bereich mit einem besonders hohen Potenzial für Big-Data-Ansätze. Neue, für die Gesundheitsforschung wissenschaftlich wie klinisch interessante Datenquellen und innovative Möglichkeiten der Datenanalyse stünden inzwischen zur Verfügung. Als hervorzuhebende Beispiele aus der Wissenschaft nennt er die Omics-Forschung und hochauflösende bildgebende Verfahren sowie aus der medizinischen Praxis die elektronische Patientenakte und frei verfügbare öffentliche Daten. Auf der Ebene der Analytik sieht er insbesondere deutliche Fortschritte bei der Informationsextraktion aus Textdaten. Der Autor verweist auch darauf, dass medizinspezifische Besonderheiten wie Datenkomplexität und rechtliche sowie ethische Rahmenbedingungen bisher zu einer weniger intensiven Anwendung von Big-Data-Ansätzen in der Medizin als in anderen Bereichen geführt haben.

*S. Bender und P. Elias* analysieren den Einfluss von gesetzlichen Entwicklungen in der Europäischen Union (EU) auf den grenzüberschreitenden Zugang zu Mikrodaten für die Forschung. Die Autoren beschreiben zwei unterschiedliche Entwicklungstendenzen: zum einen die Ambitionen, die gemeinsame Nutzung von Daten über die Staatsgrenzen zu ermöglichen, zum anderen die Bemühungen um einen innerhalb der EU harmonisierten Gesetzesrahmen, um Missbrauch persönlicher Informationen zu verhindern. Diese Entwicklungen werden mit Bezug zur Anwendung von Big Data diskutiert.

*K. Wegscheider und U. Koch-Gromus* diskutieren, inwieweit Big-Data-Ansätze für die Weiterentwicklung der Versorgungsforschung genutzt werden können. Die Autoren verweisen einerseits darauf, dass Kernaufgaben der Versorgungsforschung wie Theoriebildung, wertende Evaluation oder der Nachweis von Interventionserfolgen von der ausschließlich auf Korrelationen aufbauenden Big-Data-Methodik weder unterstützt noch ersetzt werden können. Andererseits sehen sie bei Big Data das Potenzial, die Versorgungsforschung bei ihren Aufgaben in der Datenverknüpfung, bei der Abbildung von Versorgungspfaden, beim schnellen Zugriff z. B. auf Daten zum Inanspruchnahmeverhalten sowie bei der Prädiktion und der Hypothesengenerierung substanziell zu unterstützen.

Die nachfolgenden drei Beiträge beziehen sich nicht auf Big Data, sondern auf andere Ansätze im Kontext großer Datensätze. *W. Ahrens und K.-H. Jöckel* beschreiben in ihrem Beitrag die bereits oben erwähnte Nationale Kohorte. In ihrem Rahmen ist in über 18 Studienzentren in Deutschland die Untersuchung von insgesamt 200.000 Frauen und Männern im Alter von 20–69 Jahren vorgesehen. Inhaltlich zielt die Kohortenstudie auf die Erforschung von Krankheitsursachen im Zusammenhang mit Lebensgewohnheiten, genetischen, sozioökonomischen, psychosozialen und umweltbedingten Faktoren bei den großen Volkskrankheiten. Die Autoren beschreiben ausführlich, warum die Nationale Kohorte trotz ihres außergewöhnlich großen Umfangs nicht als Big-Data-Projekt angesehen werden kann.

*S. Jacobs, C. Stallmann und I. Pigeot* verweisen darauf, dass die von ihnen grundsätzlich favorisierten Kohortenansätze fehleranfällig gegenüber Selektionseffekten oder Verzerrungen aufgrund lü-

ckenhafter Erinnerungen sein können. Diese Beeinträchtigungen der Aussagekraft von Kohortenstudien können nach ihrer Ansicht zumindest teilweise dadurch ausgeglichen werden, dass Primärdaten aus der Kohortenstudie auf Individualebene mit Sekundär- und Registerdaten (insbesondere Kranken- und Rentenversicherungsdaten, Angaben der Bundesagentur für Arbeit sowie Krebsregisterdaten) verlinkt werden. Die besonderen Stärken solcher Sekundär- und Registerdaten liegen im Umfang an Detailinformationen, in den langen Beobachtungszeiträumen und großen Populationen. Die Autorengruppe diskutiert die mit diesen Datenquellen gleichzeitig verbundenen Einschränkungen und die in Deutschland *bestehenden rechtlichen Restriktionen*.

*E. Garbe und I. Pigeot*verweisen in ihrem Beitrag auf den Nutzen großer Gesundheitsdatenbanken für die Arzneimittelrisikoforschung wegen der damit verbundenen Möglichkeit eines prospektiven Monitorings der Arzneimittelsicherheit nach Zulassung. Dieser Zugang ist aber nur für eine kleinere Teilgruppe von Arzneimittel-exponierten Patienten eine realistische Option. Als weiterführende Ansätze beschreiben und analysieren die Autorinnen verschiedene in den USA erprobte Modelle einer Verknüpfung der Daten unterschiedlicher elektronischer Gesundheitsdatenbanken (Sentinel-Initiative, Observational Medical Outcomes Partnership). Für Deutschland sehen die Autorinnen die Notwendigkeit von gesetzlichen Veränderungen, um die Datengrundlage für die Arzneimittelsicherheit zu verbessern.

Die nachfolgenden drei Beiträge setzen sich mit unterschiedlichen Nutzungsfeldern und Anwendungsmöglichkeiten von Big Data-Ansätzen auseinander. *M. Zwick* prognostiziert, dass Big Data in den kommenden Jahren nahezu alle Bereiche der amtlichen Statistik verändern wird. Es wird künftig nach Ansicht des Autors vor allem darum gehen, die richtigen, schon vorhandenen Datenbestände zu identifizieren und in adäquater Weise zu nutzen. Auf diesem Wege sind noch zahlreiche Fragen zur Qualität der Daten, zum Datenschutz und zur sicheren Verfügbarkeit der Daten zu lösen. Auf der Ebene der EU gibt es mit der Big-Data-Roadmap und dem Big-Data-Action-Plan bereits konkrete Umsetzungsstrategien zur Integration der neuen Datenbestände, die mit Big Data umschrieben werden und einen potenziellen Nutzen für die Weiterentwicklung amtlicher Statistiken aufzeigen können.

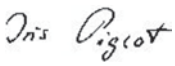
Im anschließenden Beitrag vertreten *S. Schulz und P. López García* die These, dass mit Big-Data Technologien zur computergestützten Verarbeitung menschlicher Sprache unterstützt werden können. Die Autoren schätzen die Kombination von natürlichsprachlichen Technologien, semantischen Ressourcen und Big-Data-Analytics als vielversprechend ein.

Die letzten beiden Beiträge beziehen sich auf übergreifende ethische und gesellschaftliche Aspekte des Umgangs mit großen Datensammlungen und Big Data. *E. Witt* stellt in ihrem Beitrag zunächst die wesentlichen Arten der großen Datensammlungen mit Gesundheitsbezug und deren aktuelle Anforderungen bezüglich des Datenschutzes dar. Darauf aufbauend werden Risiken und Möglichkeiten der steigenden Transparenz und Verknüpfbarkeit von Gesundheitsdaten beschrieben und künftige Herausforderungen für den Datenschutz und den Schutz der Privatsphäre formuliert.

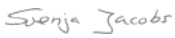
*J. Radermacher* stellt im abschließenden Beitrag einige Grundsatzüberlegungen zu Algorithmen, maschineller Intelligenz und Big Data an. Er verweist darauf, wie das Internet die Leistungsfähigkeit von Maschinen in den letzten Jahrzehnten massiv gesteigert hat. Er erwartet eine weitere Entwicklungsexplosion durch die Anwendung von Big-Data-Ansätzen. Nach Ansicht des Autors birgt eine möglicherweise unkontrolliert verlaufende Weiterentwicklung zahlreiche erhebliche Risiken für die Zivilisation. Deshalb fordert er eine „Einhegung" dieser Prozesse im Sinne einer „vernünftigen politischen Global Governance".

Wir hoffen, dass wir eine gute Abbildung des aktuellen Themas „große Datensammlungen/Big Data" durch die ausgewählten Beiträge erreicht haben. Wir wünschen den Leserinnen und Lesern eine spannende Auseinandersetzung mit den Artikeln dieses Themenhefts.

Ihre



Iris Pigeot



Svenja Jacobs

Uwe Koch-Gromus

## Korrespondenzadressen


**Prof. Dr.rer.nat. I. Pigeot**
Leibniz-Institut für Präventionsforschung und Epidemiologie – BIPS
Achterstrasse 30
28359 Bremen
pigeot@bips.uni-bremen.de

**Dr. PH S. Jacobs**
Leibniz-Institut für Präventionsforschung und Epidemiologie – BIPS
Achterstrasse 30
28359 Bremen

**Prof. Dr. med. Dr. phil. U. Koch-Gromus**
Universitätsklinikum Hamburg-Eppendorf
Martinistraße 52
20246 Hamburg



**Interessenkonflikt.** I. Pigeot, S. Jacobs und U. Koch-Gromus geben an, dass kein Interessenkonflikt besteht.