ÜBERSICHTSARBEIT

Big Data in der wissenschaftlichen Medizin – eine biostatistische Perspektive

Teil 21 der Serie zur Bewertung wissenschaftlicher Publikationen

Harald Binder, Maria Blettner

ZUSAMMENFASSUNG

Hintergrund: Durch kostengünstige Messtechniken und Speichermöglichkeiten entstehen auch in der Medizin große Datenmengen, die mit herkömmlichen Ansätzen der Datenanalyse nur schwer zu bewältigen sind. Von „Big Data" spricht man beispielsweise, wenn Datenmengen im Terabyte-Bereich (1 Terabyte = 10^{12} Byte) untersucht werden. Mit „Big Data"-Techniken wird versucht, entsprechende Datenmengen sinnvoll auszuwerten. Für die wissenschaftliche Medizin stellt sich die Frage nach dem Nutzen und der Relevanz derartiger Datensammlungen.

Methoden: Anhand von beispielhaft genannten Einsatzszenarios und einer selektiven Literaturübersicht werden Analysetechniken diskutiert und kritische Punkte aufgezeigt, die beachtet werden müssen, um Fehler im Umgang mit großen Datenmengen zu vermeiden.

Ergebnisse: Techniken des maschinellen Lernens ermöglichen es, potenziell relevante Muster zu erkennen. Allerdings müssen im Gegensatz zu herkömmlichen Analysen Anpassungen vorgenommen werden, um zum Beispiel eine Gewichtung der Patientencharakteristika vorzunehmen. Sonst würden beispielsweise zur Ähnlichkeitsbestimmung – ein Baustein vieler Verfahren – Merkmale wie Alter oder Geschlecht kein höheres Gewicht erhalten als ein einzelner von 10 000 Genexpressionswerten. Im Umgang mit den Daten können Erfahrungen aus konventionellen Beobachtungsdaten genutzt werden, um gegebenenfalls auf kausale Effekte schließen zu können.

Schlussfolgerung: Mit „Big Data"-Techniken könnten beispielsweise Beobachtungsdaten aus der Routineversorgung ausgewertet werden, wobei behandlungsrelevante Patientenuntergruppen über Clustering-Ansätze betrachtet werden. Entsprechende Auswertungen könnten klassische klinische Studien ergänzen. Durch die zunehmende Popularität von „Big Data"-Ansätzen wird eine Kombination von statistischen Techniken zur Kausalitätsanalyse in Beobachtungsdaten breiter verfügbar. Dies verspricht auch einen Gewinn für die wissenschaftliche Medizin, erfordert aber Anpassungen an die spezifischen Erfordernisse.

▶ **Zitierweise**
Binder H, Blettner M: Big data in medical science—a biostatistical view. Part 21 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2015; 112: 137–42. DOI: 10.3238/arztebl.2015.0137

Institut für Medizinische Biometrie, Epidemiologie und Informatik (IMBEI), Universitätsmedizin der Johannes-Gutenberg-Universität Mainz: Prof. Dr. oec. publ. Harald Binder; Prof. Dr. rer. nat. Maria Blettner

Ein beherrschendes Schlagwort in der Wirtschaft und Wissenschaft ist „Big Data". Damit sind wachsende Datenmengen und der Umgang mit diesen gemeint. Es ist zum Beispiel davon auszugehen, dass ein typisches Krankenhaus jährlich hunderte Terabyte (1 Terabyte = 10^{12} Byte) an Daten aus der Versorgung heraus generieren wird (1). So befindet sich zum Beispiel die Exom-Sequenzierung, die pro Patient circa fünf GigaByte (1 GByte = 10^9 Byte) an Daten erzeugt, auf dem Weg in die Routineanwendung (2). Die Analyse derartiger Datenmengen, das heißt die Organisation, die Deskription und das Ziehen von (statistisch abgesicherten) Schlüssen, ist mit traditionellen Mitteln der Informatik und Statistik kaum mehr zu bewältigen. So erfordert zum Beispiel die gemeinsame Betrachtung des Exoms mehrerer hundert Patienten ausgeklügelte informatische Ansätze und eine rechenzeitoptimierte Wahl statistischer Ansätze, um nicht an Speicherkapazitätsgrenzen zu stoßen.

Deshalb ist auch die Statistik als Disziplin gefordert, welche sich traditionell schon neben klinischen Studien auch mit Daten aus Beobachtungsstudien beschäftigt hat. Dies bedeutet unter anderem, dass Techniken mit einer Zahl von erhobenen Merkmalen pro Individuum umgehen müssen, die deutlich größer ist als die Zahl der betrachteten Individuen, wie zum Beispiel bei der Erhebung von 5 Millionen Einzelnukleotidpolymorphismen für jeden aus einer Kohorte von 100 Patienten.

Bei der folgenden Beschreibung von Einsatzszenarios, Techniken und Problemen liegt der Fokus auf der wissenschaftlichen Medizin, das heißt auf der Frage, wo und wie „Big Data"-Ansätze zur Verarbeitung großer Datenmengen zum wissenschaftlichen Erkenntnisgewinn in der Medizin beitragen können. Während die anschließende Beschreibung von korrespondierenden Datenanalysetechniken stark aus der wissenschaftlichen Perspektive heraus motiviert ist, so sollen die drei beispielhaften Szenarios auch eine bessere Orientierung im Umgang mit Routinedaten ermöglichen.

Da klinische Studien die Referenz für die vorliegende Diskussion bilden, werden Anwendungen, die sehr weit von deren Struktur entfernt liegen, nicht betrachtet, wie zum Beispiel die Vorhersage von Krankheitsausbreitung aus Suchmaschinen-Daten *(Kasten)*.

KASTEN

Debatte um das „Big Data“-Vorzeigeprojekt „Google Flu Trends“

Im Projekt „Google Flu Trends“ (3) wird aus der Häufigkeit bestimmter Anfragen an die Suchmaschine Google die Influenzaaktivität auf Regionalebene in einer großen Zahl von Ländern vorhergesagt. Die ursprüngliche Veröffentlichung (3) zeigt, dass sich so sehr genau Daten vorhersagen lassen, die traditionell deutlich aufwendiger erhoben werden müssen, zum Beispiel durch die US Centers for Disease Control and Prevention (CDC), und erst mit zeitlicher Verzögerung zur Verfügung stehen. Die potenziell schnellere Reaktionsmöglichkeit auf Basis des Google-Ansatzes wird häufig als „Big Data“-Erfolg angeführt. Allerdings zeigen spätere Untersuchungen (4), dass es nach dem in (3) betrachteten Zeitraum gravierende systematische Vorhersageabweichungen gab. Diese gehen möglicherweise darauf zurück, dass der Suchmaschinenalgorithmus aus geschäftlichen Gründen, das heißt zur Optimierung der primären Nutzung, modifiziert wurde und so die Influenzavorhersage als Sekundärnutzung in Mitleidenschaft gezogen hat.

Auch werden informatische Konzepte zur technischen Umsetzung nicht dargestellt, wie zum Beispiel „Cloud Computing“ (5). Der Fokus liegt vielmehr auf biostatistischen Aspekten, wie zum Beispiel der möglichst unverzerrten Schätzung von Therapieeffekten, die eine wesentliche Voraussetzung zum Erkenntnisgewinn in der wissenschaftlichen Medizin sind (6).

„Big Data“-Szenarios

Diagnose auf Basis hochauflösender Messungen

Schon die Verfügbarkeit von Microrarray-Techniken machte es möglich, Patienten zum Diagnosezeitpunkt auf mehreren molekularen Ebenen zu charakterisieren, zum Beispiel über Einzelnukelotidpolymorphismen, DNA-Methylierung, mRNAs oder microRNAs (7). Dies führt zu mehreren Millionen Messwerten pro Patient. Aus diesen könnten über statistische Verfahren Parameter identifiziert werden, um zwischen verschiedenen Krankheitsbildern zu unterscheiden oder Therapieentscheidungshilfen zu liefern.

Neuere Sequenzierungstechniken (bekannt als „Next Generation Sequencing“) bieten eine höhere Auflösung und steigern die Zahl der Variablen, die betrachtet werden können (8). Allerdings ist für eine kleinere Zahl von Patienten diese Menge an Daten nach Vorverarbeitungsschritten nicht mehr so groß und es sind noch keine speziellen informatischen Ansätze zur Datenhandhabung erforderlich. Beispielsweise belegt die Genexpressionsinformation zu 22 000 Genen für 400 Patienten weniger als ein GByte und kann so auf Standard-PCs verarbeitet werden. Die Information zu 5 Millionen Einzelnukleotidpolymorphismen für 400 Patienten hat ein Volumen von circa 100 GByte und kann von Computerservern, wie sie in kleineren wissenschaftlichen Arbeitsgruppen zur Verfügung stehen, im Hauptspeicher verarbeitet werden.

Tatsächliche „Big Data“-Herausforderungen entstehen dann, wenn zum Beispiel die Rohdaten von Messungen oder Messungen mehrerer molekularer Ebenen von mehreren tausend Individuen gemeinsam betrachtet werden sollen. Die Datenmenge wird hier zu einem wichtigen Faktor bei der Wahl der Analysestrategie, da sich verschiedene statistische Verfahren unterschiedlich gut auf größere Datenmengen übertragen lassen. Dies ist nicht nur in epidemiologischen Kohorten der Fall, sondern auch in einem Diagnose-Szenario, wenn die Daten eines Patienten mit externen Quellen abgeglichen werden sollen (9). So bietet zum Beispiel der Cancer Genome Atlas (TCGA) Daten mehrerer molekularer Ebenen an. Ein automatisierter Abgleich ist eine informatische und statistische Herausforderung (10).

Kontinuierliches Monitoring von gesunden Individuen

Im 100K-Projekt werden von gesunden Individuen neben einer initialen Bestimmung der Genomsequenz über einen Zeitraum von mehreren Jahren mehrmals im Jahr klassische Parameter der klinischen Chemie, Teile des Mikrobioms und organspezifische Proteine, und in einem engen zeitlichen Muster Herz-, Atmungs- und Schlafparameter erhoben (11). Seit dem Jahr 2014 findet als Testlauf eine erste Studie mit Messungen dieser Art für 108 Individuen statt, unter anderem um die technische Machbarkeit und potenzielle Datenverwendung zu evaluieren. Diesem Vorhaben liegt die Idee zugrunde, dass schon lange vor einem Diagnosezeitpunkt relevante Änderungen der Werte dieser Parameter stattgefunden haben, die bei frühzeitiger, kontinuierlicher Überwachung schon zu korrigierenden Maßnahmen führen könnten (12).

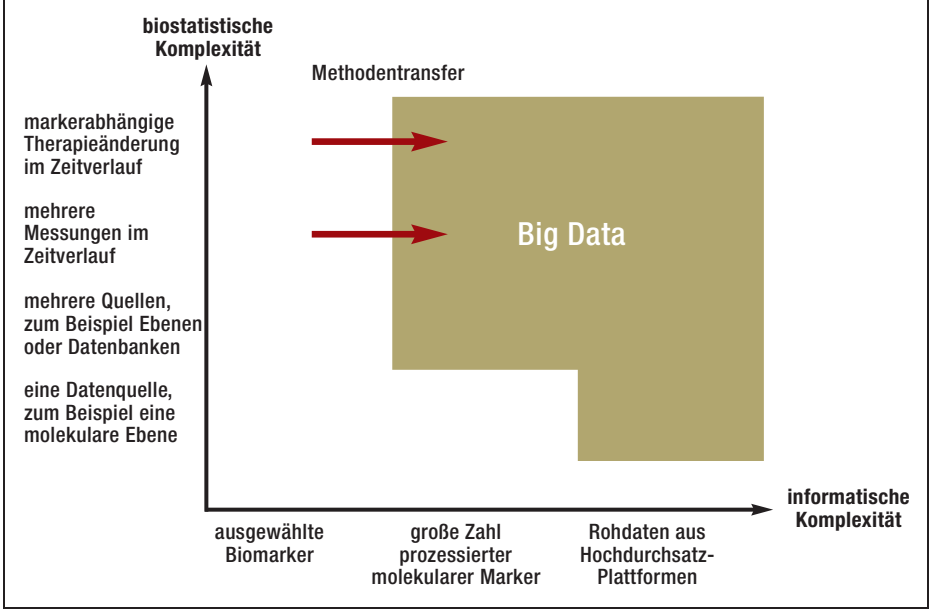
Hier kommt zu einer potenziell großen Zahl von gemessenen Parametern noch eine zeitliche Dimension hinzu. Um an zeitlichen Mustern frühzeitig problematische Entwicklungen ablesen zu können, muss die Datenanalyse eine explizite Suche nach zeitlichen Mustern in hochdimensionalen Daten vorsehen.

Die Komplexität bei kontinuierlichem Monitoring erhöht sich durch gleichzeitige Interventionen, wie zum Beispiel individuelle Ernährungsberatung. Um die Folgen von Interventionen abschätzen zu können sind ähnlich komplexe Ansätze notwendig wie bei der Nachverfolgung im Therapieprozess ab Diagnose in einem klinischen Kontext (13).

Vorhersage und Therapieentscheidung

Ein weiteres Szenario ist die Betrachtung von molekularen Charakteristika im Behandlungsverlauf. Für eine kleinere Zahl von Biomarkern geschieht dies bereits routinemäßig in klinischen Registern. So liegen zum Beispiel im Mainzer Register zum hepatozellulären Karzinom (14) für mehr als tausend Individuen Daten für teilweise mehr als ein Dutzend Messzeitpunkte vor, das heißt im Behandlungsverlauf entsteht für jedes Individuum eine umfangreichere Datenmenge.

GRAFIK



Informatische und biostatistische Komplexität von verschiedenen „Big Data"-/Datenanalyse-Fragestellungen

In darauf aufbauenden Prognose-Fragestellungen wird, ausgehend von Messungen zu einem definierten Zeitpunkt, eine Wahrscheinlichkeitsvorhersage für zukünftige Ereignisse erstellt, zum Beispiel für Metastasierung oder Todesfälle. Nach einer zukünftig zu erwartenden Ergänzung klinischer Register um hochauflösende molekulare und/oder Bildgebungsmessungen könnten zum Beispiel Krebspatienten anhand ihres Genexpressionsprofils in Gruppen mit hohem beziehungsweise niedrigerem Sterberisiko aufgeteilt werden, zum Beispiel als Basis für Therapieentscheidungen (15). Ähnlich zu Diagnose-Fragestellungen können hier Messungen von wenigen hundert Patienten zu einem einzigen Zeitpunkt nach Vorverarbeitungsschritten auf Standard-PCs verarbeitet werden (16). Die zeitliche Dimension der Messungen erhöht die Datenmenge und Komplexität deutlich.

Eine zusätzliche Herausforderung ergibt sich durch den parallel zur Messung laufenden Behandlungsprozess. Bei Patienten werden kontinuierlich Behandlungsentscheidungen getroffen, die auf den gemessenen Charakteristika basieren und diese wiederum beeinflussen. Neben dem zeitlichen Gitter der Messungen muss also gleichzeitig das zeitliche Muster der Behandlungsentscheidungen betrachtet werden, um zum Beispiel Patienten untereinander vergleichen und optimierte Therapieentscheidungen ableiten zu können. Vor allem diese Kombination bildet die Basis für eine personalisierte Medizin.

Der in den angeführten Szenarios unterschiedliche Grad an informatischer und biostatistischer Komplexität von verschiedenen „Big Data"-/Datenanalyse-Fragestellungen ist in der Grafik nochmals zusammengefasst *(Grafik)*.

Techniken

Ein Merkmal von „Big Data"-Szenarios ist es, dass die anfallenden Daten mit konventionellen Methoden nur noch schwierig zu handhaben sind. Dies betrifft als ersten Schritt der Datenanalyse die Deskription. So würde zum Beispiel bei zehn potenziellen Biomarkern typischerweise eine Mittelwertstabelle zur Deskription erstellt werden, bei 10 000 oder mehr potenziellen Markern ist eine derartige Tabelle nicht mehr hilfreich. Die Mustererkennung, das heißt die Identifikation relevanter, potenziell häufiger Muster, muss gerade bei „Big Data"-Anwendungen durch Techniken des maschinellen Lernens unterstützt werden, die automatisiert Muster erkennen und eine Verdichtung oder Vorauswahl liefern können (17).

Sogenannte „unsupervised"-Techniken suchen dabei zur Deskription zum Beispiel nach dem häufigen gemeinsamen Auftreten bestimmter Patientencharakteristika. Ein Beispiel dafür ist der „Bump Hunting"-Ansatz, der die Definitionskriterien von häufigen Gruppen von Individuen schrittweise verfeinert (18). Als Alternative dazu können Clustering-Ansätze Gruppen ähnlicher Patienten identifizieren (19). Wenn gleichzeitig zum Beispiel Biomarker identifiziert werden sollen, die in Bezug auf diese Patientengruppen ähnliche Muster zeigen, so stehen dafür Biclustering-Verfahren zur Verfügung (20).

Im Gegensatz dazu sind „supervised"-Ansätze auf ein bestimmtes Zielkriterium ausgerichtet, zum Beispiel die Vorhersage des Ein-Jahres-Überlebens auf Basis des Tumor-Genexpressionsprofils zum Diagnosezeitpunkt. Wesentlich ist hierbei die automatisierte Auswahl von wenigen Patientenmerkmalen oder zum Beispiel Genexpressionsparametern, die

TABELLE

Verschiedene Klassen von Verfahren des maschinellen Lernens mit typischem Einsatzzweck und beispielhaften Ansätzen*

	Modellfrei	Modellbasiert
unsupervised	Deskription, Mustererkennung, z. B. „bump hunting“ (18)	Verteilung von (unbekannten) Gruppen, z. B. Mischmodelle (32)
supervised	Vorhersage, z. B. „random forests“ (22)	Vorhersage, Prädiktoren identifizieren, z. B. regularisierte Regression (23)

*„unsupervised“ bedeutet dabei Mustersuche ohne ein quantifizierbares Zielkriterium (zum Beispiel Vorhersagegüte in Bezug auf den in den Daten bekannten Überlebensstatus), während bei „supervised“ ein Zielkriterium vorliegt.

gut zur Vorhersage geeignet sind. Eine weitere wichtige Unterscheidung besteht darin, ob und in welchem Umfang den jeweiligen Ansätzen ein statistisches Modell zugrunde liegt, das heißt eine mathematisch explizit spezifizierte Form des Zusammenhangs zwischen den beobachteten Größen. Modellbasierte Ansätze stammen oft aus der klassischen Statistik (siehe [21] für Erweiterungen von Regressionsmodellen), während modellfreie Ansätze oft Wurzeln in der Informatik haben (22). Prominente modellbasierte Ansätze sind regularisierte Regressionsverfahren (23) und die „Logic Regression“ (24). Bekannte modellfreie Ansätze sind „Zufallswälder“ (22) und „Support Vector Machines“ (25).

Modellbasierte Ansätze sind ähnlicher zur in klinischen Studien verwendeten Statistik. Während allerdings klinische Studien dafür ausgelegt sind, den Effekt einer Einflussgröße, typischerweise den Effekt einer Therapie, genau zu quantifizieren, das heißt unverzerrt und mit geringer Variabilität, wird die Analyse einer großen Zahl von potenziellen Einflussgrößen, zum Beispiel vieler Biomarkerkandidaten, damit erkauft, dass zwar wichtige Marker identifiziert, aber deren Effekte nicht mehr unverzerrt geschätzt werden können (26).

Bei modellbasierten Ansätzen werden die vorliegenden Daten in Form eines geschätzten Modells zusammengefasst, auf Basis dessen zum Beispiel Vorhersagen für zukünftige Patienten getroffen werden. Bei modellfreien Ansätzen findet diese Aggregation in anderer Form statt. Beim Zufallswald-Ansatz („random forrest“) wird zum Beispiel eine große Zahl von Entscheidungsbäumen (typischerweise 500 und mehr) auf jeweils zufällig leicht modifizierten Versionen der Daten gebildet (27). Für neue Patienten wird aus jedem dieser Bäume eine Vorhersage bestimmt, zum Beispiel eine Sterbewahrscheinlichkeit, und die vorhergesagten Werte werden (typischerweise durch Mittelung) kombiniert. Allerdings ist es schwierig, dabei den Einfluss einzelner Patientencharakteristika auf die Vorhersage zu beurteilen (28). Derartige modellfreie Ansätze sind damit eher für die Vorhersage als für den Erkenntnisgewinn in Bezug auf die zugrundeliegenden Zusammenhänge geeignet (27).

Eine extreme Form der modellfreien Ansätze benutzt direkt die Daten aller bisher beobachteten Individuen, um zum Beispiel Vorhersagen für neue Patientinnen und Patienten zu treffen. Als Beispiel identifizieren „Nächste-Nachbar“-Ansätze diejenigen Individuen, die den neuen Patienten am ähnlichsten sind, und sagen klinische Endpunkte anhand der Beobachtungen für diese ähnlichen Individuen vorher (29). An diese Idee angelehnte „case-based-reasoning“-Ansätze (30) entsprechen intuitiv der potenziellen ärztlichen Vorgehensweise basierend auf Erfahrungen mit bisherigen Patienten. Eine weitere Variante besteht darin, für Gruppen ähnlicher Individuen Vorhersagemodelle zu entwickeln (31). Die *Tabelle* zeigt eine Übersicht der verschiedenen Ansätzen mit beispielhaften Techniken und dem typischen Verwendungszweck.

Gerade bei großen Datenmengen ist es wichtig zu unterscheiden, ob eine Aggregation (zum Beispiel auf Basis eines Modells) vorliegt, oder ob immer die Daten aller Individuen vorgehalten werden müssen, um beispielsweise Vorhersagen für neue Fälle treffen zu können. Auch aus Datenschutzperspektive ist ein dauerhafter Zugriff auf große, potenziell verteilt gelagerte Patientendaten problematisch (33). Aus technischer Sicht ergeben sich weitere Probleme, wenn die Sammlung von Patientendaten ständig wächst und zum Beispiel die Vorhersage daher kontinuierlich aktualisiert werden soll. Für derartiges Lernen aus Datenströmen können Anpassungen entweder in regelmäßigen Intervallen durchgeführt werden, zum Beispiel mit Neuschätzung eines Regressionsmodells, oder es werden speziell angepasste Verfahren eingesetzt, bei denen die Vorhersagemodelle Individuum für Individuum anpasst werden können (34).

Besonderheiten der wissenschaftlichen Medizin

Die im letzten Abschnitt vorgestellten Ansätze wurden oft nicht für die spezifischen Anforderungen der Medizin entwickelt. Dies betrifft besonders die Berücksichtigung:

- der unterschiedlichen Arten von Patientencharakteristika
- der zeitlichen Struktur
- der Behandlungsinformation.

Ohne spezielle Anpassung behandeln Verfahren des maschinellen Lernens, das heißt Verfahren, die automatisiert Muster erkennen und eine Verdichtung oder Vorauswahl liefern können, alle Messungen oder Patientencharakteristika in gleicher Art und Weise. So würden zum Beispiel zur Ähnlichkeitsbestimmung, die ein Baustein vieler Verfahren ist, Charakteristika wie Alter oder Geschlecht kein höheres Gewicht bekommen als jeder einzelne von 20 000 gemessenen Genexpressionswerten. Schon allein zur Optimierung der Vorhersageleistung ist aber eine Unterscheidung zwischen klinischen Merkmalen und weiteren Charakteristika, zum Beispiel hochdimensionalen molekularen Messungen, vorteilhaft (35).

Beim kontinuierlichen Monitoring von Individuen und bei der Betrachtung von Messungen im zeitlichen Verlauf der Behandlung liegt neben der potenziell hohen Dimension der Messungen aufgrund der Zeitstruktur eine zusätzliche Dimension vor, die bei der Datenanalyse berücksichtigt werden muss (36). So ist zum Beispiel der Diagnosezeitpunkt eine wichtige Referenz, wenn spätere molekulare Messungen zwischen Individuen verglichen beziehungsweise deren Ähnlichkeit im Rahmen eines Verfahrens des maschinellen Lernens bestimmt werden soll. Zusätzliche Komplikationen treten hier durch unterschiedlich lange Nachverfolgungszeiträume für verschiedene Individuen auf. Dies entspricht der Zensierungsproblematik, wie sie in klinischen Studien für die Betrachtung des interessierenden Endpunktes durch Verfahren wie den Kaplan-Meier-Schätzer oder die Cox-Regression angegangen wird. Gerade Verfahren des maschinellen Lernens müssen für derartige Zeitstrukturen erst speziell angepasst werden. Eine vereinfachende Reduktion, zum Beispiel auf einen binären Endpunkt trotz Zensierung, kann zu deutlich verzerrten Ergebnissen führen (21). Auch ohne Zensierung kann ein unregelmäßiges Gitter an Messzeitpunkten, wie es oft bedingt durch die klinische Routine vorliegt, zu Verzerrungen führen (37).

Schließlich kommt der Therapieinformation und den Zeitpunkten der Therapieentscheidung und Änderung eine wesentliche Rolle bei der Suche nach Mustern in potenziell großen Datenmengen zu. In der klinischen Routine wird die Therapieentscheidung einerseits durch Messwerte beeinflusst, aber auch die Therapieentscheidung wird (zukünftige) Messwerte beeinflussen. Wenn in einer derartigen Konstellation beispielsweise der Effekt einer Therapie auf das Überleben bestimmt werden soll und zur Vergleichbarmachung von Patienten auf einen im Zeitverlauf gemessenen Laborparameter adjustiert wird, typischerweise durch Adjustierung in einem Regressionsmodell, so kann diese Adjustierung einen Teil des Therapieeffektes, der wiederum am Laborparameter abzulesen ist, verdecken. Allgemein wird diese Problematik, die zu verzerrten Schätzung von Therapieeffekten in jegliche Richtung führen kann, als „time-dependent confounding“ bezeichnet (38).

Für klassische biostatistische Analysen von Beobachtungsdaten wurden Verfahren entwickelt, die mit zensierten Beobachtungen umgehen können, und auch Ansätze zur gemeinsamen Betrachtung kontinuierlich erhobener Messungen und eines potenziell zensierten klinischen Endpunktes (39). Ebenso gibt es verschiedene Ansätze zum Umgang mit der „time-dependent confounding“-Problematik. Während derartige Ansätze bisher kaum mit Verfahren des maschinellen Lernens kombiniert wurden, so besteht doch prinzipiell das Potenzial dafür. So basiert beispielsweise der „sequential Cox“-Ansatz für die „time-dependent confounding“-Problematik auf umgeformten Daten (40), auf die auch Verfahren des maschinellen Lernens angewendet werden könnten.

Diskussion

Der Begriff „Big Data“ umspannt verschiedenste Disziplinen, Anwendungen und vielfältige statistische und informatische Ansätze.

Für Anwendungen der wissenschaftlichen Medizin muss dabei, wie aufgezeigt, die Unterschiedlichkeit verschiedener Patientencharakteristika, die Zeitstruktur und die Behandlungsinformation berücksichtigt werden. Während es bereits einige Ansätze des maschinellen Lernens gibt, die manchen dieser Anforderungen Rechnung tragen und deshalb auch für „Big Data“-Anwendungen in diesem Bereich eingesetzt werden könnten, so gibt es noch ein großes Potenzial für die Entwicklung adäquater Ansätze zur automatisierten Mustererkennung. Viele dieser noch zu entwickelnden Ansätze werden voraussichtlich auch für Anwendungen nützlich sein, die auf einem Kontinuum der Komplexität gerade noch nicht als „Big Data“-Problem angesehen werden. Damit wird möglicherweise grundsätzlich die Verwendung von Beobachtungs- und vor allem Routinedaten erleichtert. Zwar werden im Vergleich zu klinischen Studien nur sehr schwer verzerrungsfreie Ergebnisse mit Hilfe von „Big Data“-Ansätzen erzielt werden können, doch versprechen letztere zumindest eine wertvolle Ergänzung für den Gewinn an medizinischer Erkenntnis.

KERNAUSSAGEN

- Auch die wissenschaftliche Medizin ist mit „Big Data“-Problemen konfrontiert, insbesondere wenn molekulare Messungen auf mehreren Ebenen oder Routinedaten mit kontinuierlichem Monitoring betrachtet werden sollen.
- Automatisierte Mustererkennung, zum Beispiel über Clustering, kann in großen Datenmengen die Rolle der Deskription übernehmen, wie sie traditionell als erster Schritt der statistischen Analyse durchgeführt wird.
- Als Besonderheit der Medizin muss bei Verfahren zur Datenanalyse die Gewichtung einzelner Patientenmerkmale berücksichtigt werden, zum Beispiel Alter und Geschlecht in Relation zu Tausenden von Genexpressionsmesswerten.
- Zur Analyse von Kausalität aus gemeinsam erhobener Therapieinformation und molekularen Markern muss die zeitliche Abfolge berücksichtigt werden, zum Beispiel durch Adaptation von existierenden Verfahren für Beobachtungsdaten.
- Durch die zunehmende Popularität von „Big Data“-Ansätzen werden korrespondierende Datenanalysetechniken breiter verfügbar, was einen Gewinn für die wissenschaftliche Medizin verspricht.

Interessenkonflikt
Die Autoren erklären, dass kein Interessenkonflikt besteht.

Manuskriptdaten
eingereicht: 8. 5. 2014, revidierte Fassung angenommen: 18. 11. 2014

LITERATUR

1. Sejdic E: Adapt current tools for handling big data (Correspondence). Nature 2014; 507: 306.
2. Tripathy D, Harnden K, Blackwell K, Robson M: Next generation sequencing and tumor mutation profiling: Are we ready for routine use in the oncology clinic? BMC Med 2014; 12: 140.
3. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L: Detecting influenza epidemics using search engine query data. Nature 2009; 457: 1012–4.
4. Lazer D, Kennedy R, King G, Vespignani A: The parable of google flu: Traps in big data analysis. Science 2014; 343: 1203–5.
5. Marx V: The big challenges of big data. Nature 2013; 498: 255–60.
6. Chiolero A: Big data in epidemiology. Epidemiology 2013; 26: 938–9.
7. Cho YJJ, Tsherniak A, Tamayo P, et al.: Integrative genomic analysis of medulloblastoma identifies a molecular subgroup that drives poor clinical outcome. J Clin Oncol 2011; 29: 1424–30.
8. Marioni J, Mason C, Mane S, Stephens M, Gilad Y: RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genom Res 2008; 18: 1509–17.
9. Huerta M, Munyi M, Expósito D, Querol E, Cedano J: MGDB: crossing the marker genes of a user microarray with a database of public-microarrays marker genes. Bioinformatics 2014; 30: 1780–1.
10. Robbins DE, Grüneberg A, Deus HF, Tanik MM, Almeida JS: A self-updating road map of the cancer genome atlas. Bioinformatics 2013; 29: 1333–40.
11. Hood L, Price ND: Demystifying disease, democratizing health care. Sci Transl Med 2014; 5: 225.
12. Hood L, Friend SH: Predictive, personalized, preventive, participatory (P4) cancer medicine. Nat Rev Clin Oncol 2011; 8: 184–7.
13. Gibbs WW: Medicine gets up close and personal. Nature 2014; 506: 144.
14. Weinmann A, Koch S, Niederle IM, Schulze-Bergkamen H, et al.: Trends in epidemiology, treatment and survival of hepatocellular carcinoma patients between 1998 and 2009: an analysis of 1066 cases of a German HCC registry. J Clin Gastroenterol 2014; 48: 279–89.
15. Simon R: Roadmap for developing and validating therapeutically relevant genomic classifiers. J Clin Oncol 2005; 23: 7332–41.
16. Horn JDV, Toga AW: Human neuroimaging as a big data science. Brain Imaging Behav 2013; 2: 323–31.
17. James G, Witten D, Hastie T, Tibshirani R: An introduction to statistical learning. New York: Springer 2013.
18. Friedman JH, Fisher NI: Bump hunting in high-dimensional data. Stat Comput 1999; 9: 123–43.
19. Andreopoulos B, An A, Wang X, Schroeder M: A roadmap of clustering algorithms: finding a match for a biomedical application. Brief Bioinform 2009; 10: 297–314.
20. Eren K, Deveci M, Küçüktunç O, Çatalyürek ÜV: A comparative analysis of biclustering algorithms for gene expression data. Brief Bioinform 2013; 14: 279–92.
21. Binder H, Porzelius C, Schumacher M: An overview of techniques for linking high-dimensional molecular data to time-to-event endpoints by risk prediction models. Biom J 2011; 53: 170–89.
22. Breiman L: Random Forests. Mach Learn 2001; 45: 5–32.
23. Witten DM, Tibshirani R: Survival analysis with high-dimensional covariates. Stat Methods Med Res 2010; 19: 29–51.
24. Ruczinski I, Kooperberg C, LeBlanc M: Logic Regression. J Comput Graph Stat 2003; 12: 475–511.
25. Evers L, Messow CM: Sparse kernel methods for high-dimensional survival data. Bioinformatics 2008; 24: 1632–8.
26. Porzelius C, Schumacher M, Binder H: Sparse regression techniques in low-dimensional survival settings. Stat Comput 2010; 20: 151–63.
27. Breiman L: Statistical modeling: The two cultures. Stat Sci 2001; 16: 199–231.
28. Boulesteix ALL, Janitza S, Kruppa J, König IR: Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdiscip Rev Data Min Knowl Discov 2012; 2: 493–507.
29. Kruppa J, Liu Y, Biau G, et al.: Probability estimation with machine learning methods for dichotomous and multicategory outcome: theory. Biom J 2014; 56: 534–63.
30. Glez-Peña D, Díaz F, Hernández JM, Corchado JM, Fdez-Riverola F: geneCBR: a translational tool for multiple-microarray analysis and integrative information retrieval for aiding diagnosis in cancer research. BMC Bioinformatics 2009; 10: 187.
31. Binder H, Müller T, Schwender H, et al.: Cluster-localized sparse logistic regression for SNP data. Statl Appl Genet Mol 2012; 11: 4.
32. Reich BJ, Bondell HD: A spatial dirichlet process mixture model for clustering population genetics data. Biometrics 2010; 67: 381–90.
33. Toh S, Platt R: Is size the next big thing in epidemiology? Epidemiology 2013; 24: 349–51.
34. Gaber MM, Zaslavsky A, Krishnaswamy S: Mining data streams: a review. ACM Sigmod Record 2005; 34: 18–26.
35. Binder H, Schumacher M: Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. BMC Bioinformatics 2008; 9: 14.
36. Aalen Røysland O, Gran JM, Ledergerber B: Causality, mediation and time: a dynamic viewpoint. J R Stat Soc A 2012; 175: 831–61.
37. Andersen PK, Liest K: Attenuation caused by infrequently updated covariates in survival analysis. Biostatistics 2003; 4: 633–49.
38. Daniel RM, Cousens SN, De Stavola BL, Kenward MG, Sterne JAC: Methods for dealing with time-dependent confounding. Stat Med 2012; 32: 1584–618.
39. Ibrahim JG, Chu H, Chen LM: Basic concepts and methods for joint models of longitudinal and survival data. J Clin Oncol 2010; 28: 2796–801.
40. Gran JM, Røysland K, Wolbers M, et al.: A sequential Cox approach for estimating the causal effect of treatment in the presence of time-dependent confounding applied to data from the Swiss HIV Cohort Study. Stat Med 2010; 29: 2757–68.

Anschrift für die Verfasser
Prof. Dr. oec. pub. Harald Binder
Institut für Medizinische Biometrie, Epidemiologie und Informatik (IMBEI) der Universitätsmedizin der Johannes-Gutenberg-Universität Mainz
Obere Zahlbacher Straße 69, 55101 Mainz
binderh@uni-mainz.de

Zitierweise
Binder H, Blettner M: Big data in medical science—a biostatistical view. Part 21 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2015; 112: 137–42. DOI: 10.3238/arztebl.2015.0137

@ The English version of this article is available online: www.aerzteblatt-international.de