Carolin Herrmann · Ursula Berger
Christel Weiß · Iris Burkholder
Geraldine Rauch · Jochen Kruppa *Hrsg.*

Zeig mir Health Data Science!

Ideen und Material für guten Biometrie-Unterricht mit datenwissenschaftlichem Fokus

MOREMEDIA

Springer Spektrum

Zeig mir Health Data Science!

Carolin Herrmann · Ursula Berger · Christel Weiß ·
Iris Burkholder · Geraldine Rauch ·
Jochen Kruppa (Hrsg.)

Zeig mir Health Data Science!

Ideen und Material für guten Biometrie-Unterricht mit datenwissenschaftlichem Fokus

Springer Spektrum

Hrsg.
Carolin Herrmann
Institut für Biometrie und Klinische Epidemiologie, Charité – Universitätsmedizin Berlin
Berlin, Deutschland

Ursula Berger
Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie (IBE), Ludwig-Maximilians-Universität München
München, Deutschland

Christel Weiß
Abteilung für Medizinische Statistik und Biomathematik, Medizinische Fakultät Mannheim der Universität Heidelberg
Mannheim, Deutschland

Iris Burkholder
Department Gesundheit und Pflege, Hochschule für Technik und Wirtschaft Saarbrücken, Deutschland

Geraldine Rauch
Institut für Biometrie und Klinische Epidemiologie, Charité – Universitätsmedizin Berlin
Berlin, Deutschland

Jochen Kruppa
Institut für Biometrie und Klinische Epidemiologie, Charité – Universitätsmedizin Berlin
Berlin, Deutschland

ISBN 978-3-662-62192-9 ISBN 978-3-662-62193-6 (eBook)
https://doi.org/10.1007/978-3-662-62193-6

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Planung/Lektorat: Iris Ruhmann
Springer Spektrum ist ein Imprint der eingetragenen Gesellschaft Springer-Verlag GmbH, DE und ist ein Teil von Springer Nature.
Die Anschrift der Gesellschaft ist: Heidelberger Platz 3, 14197 Berlin, Germany

Vorwort

Liebe Leserinnen und liebe Leser,

nach bereits zwei erfolgreichen Büchern mit Lehrmaterial für die Biostatistik haben Sie nun den dritten Band vor sich. Dieses Mal ist es eine erweiterte Beitragssammlung geworden. Es sind nicht nur Beiträge aus der Biostatistik enthalten, sondern aus diversen Fachbereichen, die sich mit Data Science in der Medizin beschäftigen.

Da sich (Health) Data Science über die letzten Jahre rasant zu einem Buzzword in der medizinischen Wissenschaft entwickelt hat, hat die Arbeitsgruppe Lehre und Didaktik der Biometrie als gemeinsame Arbeitsgruppe (AG) der Internationalen Biometrischen Gesellschaft, Deutsche Region und der GMDS den Lehrepreis 2020 nun in Health Data Science ausgeschrieben.

Die AG Lehre und Didaktik macht es sich zur Aufgabe, qualitativ hochwertige und abwechslungsreiche Biostatistik-Lehre an Universitäten und Fachhochschulen zu fördern und weiterzuentwickeln. Neben einem regen Austausch und der Vernetzung diverser Hochschullehrenden sowie der Weiterentwicklung von Lehrkonzepten, engagiert sich die Arbeitsgruppe auch in der Nachwuchsförderung an Schulen.

Für den Lehrepreis 2020 in Health Data Science konnten jegliche Beiträge eingereicht werden, die die Lehre in diesem Themengebiet bereichern. Dies umfasst beispielsweise Ideen für Gruppenarbeiten, Vorschläge für Übungs- und Prüfungsaufgaben und Software-Anwendungen. Außerdem sollte das Lehrmaterial gebrauchsfertig und mit Veröffentlichung frei zugänglich vorliegen. Mit der Preisausschreibung waren dieses Mal neben der Biostatistik auch ausdrücklich weitere Fachbereiche wie die Epidemiologie, Medizinische Informatik, Public Health etc. angesprochen.

Da für den Begriff Health Data Science noch keine eindeutige Definition vorliegt, haben wir führende FachvertreterInnen angesprochen, ihre Auffassung zu Health Data Science mit uns in kurzen Beiträgen zu teilen und Ihnen als erweitertes Vorwort angehängt.

Diese Beiträge sowie die Lehrmaterial-Beiträge der HerausgeberInnen dieses Buches waren außer Konkurrenz für die Vergabe des diesjährigen Lehrepreises. Es sind zahlreiche Beiträge mit verschiedensten Ideen und Methoden für die Lehre eingegangen: Von

Lehrmöglichkeiten zu personalisierter Medizin, über Humor in der Lehre bis hin zum effektiven R Training.

Die Jury, bestehend aus Prof. Dr. Geraldine Rauch, Prof. Dr. Christel Weiß, Prof. Dr. Iris Burkholder, Dr. Jochen Kruppa, Dr. Ursula Berger und Carolin Herrmann, hat den Lehrepreis Health Data Science 2020 an Antonia Zapf und Sinan Cevirme für ihren überzeugenden Beitrag zu Audio Response Systemen vergeben. Auf Platz zwei kam der Beitrag von Annette Aigner und auf Platz 3 der Beitrag von Stefan Englert, Greg Cicconetti und William Henner.

Zusätzliches Lehrmaterial zu den Beiträgen dieses Buches finden Sie unter https://link.springer.com/book/10.1007/978-3-662-62193-6. An dieser Stelle wollen wir uns herzlich bei Dr. Uwe Schöneberg und Christine Krüger bedanken, die bei der Formatierung und Vorab-Lektorat der Beiträge eine sehr große Hilfe waren. Außerdem gilt unser Dank der GMDS, welche den Pokal und die Preise finanziell ermöglicht haben.

Wir wünschen Ihnen nun viel Freude beim Lesen und der Planung Ihrer nächsten Lehreinheiten.

Berlin
31. Juli 2020

Carolin Herrmann
Ursula Berger
Christel Weiß
Iris Burkholder
Geraldine Rauch
Jochen Kruppa

Was ist Data Science?

Im Folgenden präsentieren führende Fachvertreter Ihre Sichtweisen auf Data Science.

Interoperable Daten als Grundlage für Data Science in der Medizin

Prof. Dr. Sylvia Thun forscht zu Themen rund um Medizininformatik, Digital Health und Interoperabilität. Sie ist Charité Visiting Professor und leitet die Core Unit „eHealth und Interoperabilität" am Berlin Institute of Health (BIH).
Dr. Moritz Lehne ist Data Scientist an der Core Unit „eHealth und Interoperabilität" des BIH und hat langjährige Erfahrung mit der Analyse von Gesundheitsdaten.

Data Science ist „in". Begriffe wie „Big Data", „Maschinelles Lernen" oder „Neuronale Netze", mit denen noch vor wenigen Jahren nur ein kleiner Kreis von Fachleuten etwas anfangen konnte, sind mittlerweile allgegenwärtig. Und mehr noch: Unternehmen und Organisationen, die sich heute *nicht* mit Data Science beschäftigen, werden von datengetriebenen Organisationen – Amazon, Apple, Google, Netflix usw. – zunehmend abgehängt. Wer Data Science kann, ist klar im Vorteil.

Auch in der Medizin spielt Data Science eine immer größere Rolle: Chatbots helfen bei der Diagnosestellung, epidemiologische Daten werden tagesaktuell in komplexen Visualisierungen bereitgestellt (z. B. während der Covid-19-Pandemie) und künstliche neuronale Netze erkennen Hautkrebs auf dermatologischen Bildern – teilweise zuverlässiger als Ärztinnen und Ärzte. Data-Science-Methoden haben großes Potenzial, die Medizin zu revolutionieren und die Gesundheitsversorgung zu verbessern (Topol 2019).

Doch was ist mit Data Science eigentlich genau gemeint? Wie aus den anderen Beiträgen dieses Vorworts ersichtlich, gibt es dazu die unterschiedlichsten Perspektiven. Mit den meisten Experten könnte man sich aber wahrscheinlich auf etwa folgende Definition einigen: „Data Science bedient sich Methoden der Statistik und Informatik,

um Erkenntnisse und Wissen aus Daten zu generieren.“ Auch wenn man behaupten kann, dass mit dem Begriff „Data Science“ damit nur alter Wein in neuen Schläuchen verkauft wird – schließlich existieren Fachgebiete wie die moderne Statistik, Informatik und selbst vermeintlich neue Methoden wie die Künstliche Intelligenz schon seit mindestens Jahrzehnten – so hat Data Science in den letzten Jahren zweifelsohne an Bedeutung gewonnen. Dies hat vor allem zwei Gründe: 1. die steigende Rechenleistung und Speicherkapazität moderner Computer; 2. die zunehmende Verfügbarkeit großer Mengen digitaler Daten, kurz „Big Data“. Die schnelleren Computer ermöglichen es, immer komplexere statistische Modelle zu berechnen; die digitalen Daten liefern den Input für diese Berechnungen.

Gerade der letzte Punkt, die Verfügbarkeit digitaler Daten, ist entscheidend. Eigentlich ist es selbstverständlich: Für Data Science braucht man gute (und viele) Daten. Dieser Aspekt wird allerdings häufig vernachlässigt und kann gerade in der Medizin einen Engpass darstellen (Lehne et al. 2019). Während in anderen Bereichen oft Terabyte digitaler Daten für Analysen zur Verfügung stehen, ist die Digitalisierung in der Medizin an vielen Stellen noch ausbaufähig (wahrscheinlich wären Faxgeräte heute in Deutschland ungefähr so verbreitet wie Musikkassetten, Diskettenlaufwerke oder Röhrenbildschirme, wenn sie im Gesundheitswesen nicht immer noch ein Standard-Kommunikationsmittel wären). In der Core Unit „eHealth und Interoperabilität“ des Berlin Institute of Health (BIH) beschäftigen wir uns daher damit, medizinische Daten in eine Form zu bringen, die Data Science mit modernen digitalen Technologien ermöglicht. Konkret befassen wir uns mit Dateninteroperabilität, d. h. der Fähigkeit, Daten über verschiedene Systeme auszutauschen und sinnvoll weiterzuverarbeiten. Erst durch interoperable Daten können die Möglichkeiten von Data Science voll ausgeschöpft werden.

Wieso ist Interoperabilität wichtig für Data Science? Ohne strukturierte Formate und eine einheitliche Sprache sind Daten schwer zu verarbeiten. Dies gilt insbesondere für die automatische Verarbeitung mit modernen Algorithmen, z. B. im Bereich der Künstlichen Intelligenz. So ist es für einen Algorithmus schwer zu erkennen, dass mit den Bezeichnungen „Herzinfarkt“, „akuter Myokardinfarkt“ oder einfach „aMI“ dasselbe medizinische Konzept gemeint ist. Die größtenteils unstrukturierten medizinischen Daten, die überdies über unzählige, meist proprietäre IT-Systeme verteilt sind, stellen daher keine gute Ausgangslage für Data Science in der Medizin dar. Denn je unstrukturierter die Daten, desto fehleranfälliger die Datenanalysen. Zwar gibt es auch Verfahren zur Verarbeitung unstrukturierter Daten (beispielsweise mit Methoden des Natural Language Processing auf unstrukturierten Textdaten) – aber kann man wirklich sicher sein, dass eine Patientin, in deren medizinischen Dokumenten das Wort „Diabetes“ auftaucht, auch wirklich unter Diabetes leidet (vielleicht ist in dem Dokument nur vermerkt, dass es in der Verwandtschaft der Patientin Diabetesfälle gab)? Eine möglichst strukturierte Beschreibung medizinischer Daten ist daher unabdingbar. Interoperabilität stellt sicher, dass Daten syntaktisch und semantisch eindeutig definiert sind, d. h. dass sie ein einheitliches Format und eindeutige Bezeichnungen haben.

Um Daten interoperabel zu machen und sie damit system-, institutions- und länderübergreifend verarbeiten zu können, ist internationale Zusammenarbeit erforderlich. In der Medizin ist hier die Arbeit internationaler Standardisierungsorganisationen wie Health Level 7 (HL7) oder Integrating the Healthcare Enterprise (IHE) besonders wichtig. Diese Organisationen entwickeln in transparenten Prozessen einheitliche Datenformate und -strukturen zum Austausch medizinischer Informationen, wie z. B. der zunehmend an Bedeutung gewinnende Standard „Fast Healthcare Interoperability Resources“ (FHIR) von HL7. Zur einheitlichen Benennung medizinischer Konzepte sind darüber hinaus internationale Terminologien und Nomenklaturen erforderlich – also einheitliche Vokabularien, die sicherstellen, dass unterschiedliche Systeme dieselbe Sprache sprechen. Die umfassendste dieser Nomenklaturen in der Medizin ist SNOMED CT mit aktuell über 350.000 Konzepten aus den unterschiedlichsten Bereichen der Medizin und Gesundheitsversorgung. Für das oben genannte Beispiel des Herzinfarkts gibt es hier das Konzept „Myocardial Infarction“ mit einer eindeutigen Nummer, so dass dieser medizinische Sachverhalt eindeutig benannt werden kann – unabhängig von der verwendeten Sprache („Herzinfarkt“, „infarto de miocardio“ usw.) oder eventueller Synonyme („Heart Attack“, „Cardiac Infarction“). Die Konzepte der Nomenklatur stehen außerdem in komplexen Beziehungen und Hierarchien zueinander, die ebenfalls genau definiert sind (beispielsweise ist der Herzinfarkt hierarchisch unter dem Konzept der Herzerkrankungen angeordnet).

Um die digitale Daten optimal für Data Science zu nutzen, ist es außerdem wichtig, dass Daten geteilt und wiederverwendet werden können (beispielsweise Daten aus wissenschaftlichen Studien). Hier sind die sogenannten FAIR-Prinzipien hilfreich. FAIR steht für „findable“, „accessible“, „interoperable“ und „reusable“, d. h. Daten müssen auffindbar, zugänglich, interoperabel und wiederverwendbar sein (Wilkinson et al. 2016). Dies erfordert Meta-Daten, d. h. zusätzliche, beschreibende Daten, die Auskunft über einen Datensatz geben. Erst dadurch können existierende Datenbestände von Data Scientists gefunden, hinsichtlich Inhalt und Qualität beurteilt und für Analysen verwendet werden.

Aus unserer Sicht beinhaltet Data Science daher nicht nur die Anwendung von Methoden zur Datenanalyse, sondern auch die Schaffung digitaler Infrastrukturen, die die Anwendung dieser Methoden überhaupt erst ermöglichen. Denn selbst die besten Algorithmen sind nutzlos, wenn sie keinen Zugriff auf digitale Daten haben oder diese nicht sinnvoll interpretieren können. Unsere Erfahrungen aus der Medizin zeigen, dass erst durch interoperable Daten das Potenzial von Data Science optimal genutzt werden kann. Eine Verbesserung der Dateninteroperabilität – mit einheitlichen Formaten und Vokabularien und unter Einhaltung der FAIR-Prinzipien – ist daher ein wichtiger Aspekt von Data Science.

Literatur

Lehne M, Sass J, Essenwanger A, Schepers J, Thun S (2019) Why digital medicine depends on interoperability. NPJ Digital Med 2(79). https://doi.org/10.1038/s41746-019-0158-1

Topol E (2019) Deep medicine: How artificial intelligence can make healthcare human again. Basic Books, New York

Wilkinson M, Dumontier M, Aalbersberg I et al (2016) The FAIR guiding principles for scientific data management and stewardship. Sci Data 3. https://doi.org/10.1038/sdata.2016.18

Methoden der Data Science in der Bioinformatik

Prof. Dr. Klaus Jung hat Statistik an der Universität Dortmund studiert und bereits mit seiner Diplomarbeit über die Analyse von Gen- und Proteinexpressionsdaten den Schwenk zur Bioinformatik gewagt. Nach der Promotion in Dortmund und verschiedenen Postdoc-Stationen, u. a. am Institut für Medizinische Statistik in Göttingen, wurde er 2015 zum Professur für Genomics and Bioinformatics of Infectious Diseases an die Tierärztliche Hochschule Hannover berufen. Dort erforscht er Algorithmen für die Analyse molekularbiologischer Hochdurchsatzdaten, u. a. unter Verwendung von Verfahren des Maschinellen Lernens, der Meta-Analyse und Evidenzsynthese sowie der Metagenomik.

Gerade solche Fächer die sich mit Erhebung, Verarbeitung, Management und Auswertung von Daten befassen (z. B. Statistik, Biometrie, Informatik, Epidemiologie, Ökonometrie, Medizinische Informatik, Bioinformatik) sorgen bei Außenstehenden gerne für Unklarheit was ihre jeweilige Ausrichtung betrifft. Aber auch direkte Vertreter dieser Fächer können sich in längeren Diskussion darüber ergehen, wo das eine Fach aufhört und wo das andere Fach anfängt, inklusive Debatten darüber welches Fach als Urheber für bedeutende Methoden zu sehen ist.

Der Begriff „Data Science“ (feminin) ist in einer Reihe mit den oben genannten Fächern, je nach Quelle, nicht zwingenderweise der jüngste, erfährt jedoch seit einigen Jahren eine gewisse Beliebtheit, z. B. bei der Bezeichnung von Studiengängen oder bei Ausschreibungen von Arbeitsstellen. Jüngst äußerte ein Biologe mir gegenüber er sei ja auch Datenwissenschaftler, worin er nicht ganz unrecht hat, da die meisten Naturwissenschaftler Daten erfassen und auswerten.

Im Folgenden möchte ich zum einen versuchen, eine Abgrenzung zwischen der Data Science und anderen Fächern vorzunehmen, wenn auch die Übergänge fließend sind. Zum anderen werde ich einige Methoden der Data Science welche in der Bioinformatik verwendet werden vorstellen.

Die meisten Beschreibungen die über die Data Science vorliegen betrachten es als deren Aufgabe, Informationen aus bestehenden Datenbanken zu extrahieren (z. B. umfangreiche Bestände an Kundendaten in großen Unternehmen). Dies steht in klarer Abgrenzung zur Statistik und Biometrie, die sich explizit auch mit der Planung und Erhebung von Daten befassen. Es wäre jedoch falsch daraus abzuleiten, dass die Data Science nur ein Teilgebiet der Statistik ist, denn während sich die klassische Statistik in der Regel mit Daten in Tabellen- bzw. Matrix-Format befasst, schließt die Data Science auch andere Daten- und Dateitypen wie z. B. Bild- Ton- oder Textdateien mit ein. Während Statistik und Data Science also zur Erhebung und/oder Auswertung von Daten beitragen, liefert die Informatik Methoden zu Struktur und Management von Datenbanken. Allerdings wird das Datenmanagement selbst häufig auch als Aufgabe eines Data Scientists betrachtet.

Sowohl aus der Informatik als auch aus der Statistik stammen Methoden, die die Data Science zur explorativen Extraktion von Erkenntnissen und Mustern aus Datenbeständen verwendet. In der Statistik etwa wurden verschiedene, dimensionsreduzierende Verfahren wie z. B. die Hauptkomponentenanalyse entwickelt, welche es erlauben, höherdimensionale Daten in zwei- oder dreidimensionalen Abbildungen darzustellen, und damit Gruppenstrukturen und Ausreißer zu erkennen (Kruppa und Jung 2017). Weitere Verfahren wie z. B. die Clusteranalyse und solche zur Mustererkennung mittels unüberwachtem Lernen wurden von der Informatik selbst entwickelt oder weiterentwickelt.

Schließlich werden sowohl im Data Science als auch in den anderen datenbezogenen Fächern Methoden des überwachten Lernens verwendet (Diskriminanzanalyse, Support-Vector Machines, Künstliche Neuronale Netze, etc.), um auf Trainingsdaten Klassifikationsmodelle anzupassen, welche dann für die Zuordnung neuer Beobachtungseinheiten oder für Vorhersagen verwendet werden können.

Insbesondere im Bereich der Molekularbiologie sind in den letzten drei Jahrzenten große, häufig über das Internet frei zugängliche Datenbanken entstanden. Bioinformatiker, die diese Datenbanken pflegen und daraus neue Kenntnisse gewinnen, würden sich wohl in den wenigsten Fällen als Data Scientist bezeichnen. Ein Großteil der von Bioinformatikern verwendeten Methoden überlappt aber mit den Methoden der Data Science.

In Ihren Anfängen, so etwa in den 1970er und 1980erJahren, hat die Bioinformatik schwerpunktmäßig Sequenzdaten (z. B. DNA- und Aminosäuresequenzen) betrachtet. Zur Analyse dieser Daten wurden und werden immer noch Clusterverfahren verwendet, um Ähnlichkeiten zwischen verschiedenen Spezies oder Genen zu analysieren. Des weiteren kommen Klassifikationsverfahren zum Einsatz, um z. B. aus DNA-Sequenzen Proteinstrukturen hervorzusagen. In diese zeitliche Epoche fallen auch die ersten, kleineren Datenbanken mit Sequenzinformationen für Mikroorganismen.

In den 1980er Jahren wurde außerdem die 2-dimensionale Gelelektrophorese entwickelt, mit deren Hilfe die Expression vieler Proteine gleichzeitig gemessen werden kann. Als Ergebnis liegen hochdimensionale Datenmatrizen vor, welche mit den oben genannten dimensionsreduzierenden Verfahren analysiert werden können. Ähnliche

Datenmatrizen werden mit DNA Microarrays, verfügbar seit 1995, generiert. Mit diesen Arrays können, ebenfalls hochdimensionale Genexpressionsprofile gemessen werden. Für derartige Gen- und Proteinexpressiondaten wurden bald überwachte Lernverfahren verwendet und weiterentwickelt, um z. B. Diagnosen und Prognosen von Patienten zu verbessern. Da die meisten wissenschaftlichen Journale von Ihren Autoren fordern, dass sie ihre Expressionsprofile in öffentlichen Datenbanken hochladen, stieg in den letzten beiden Jahrzehnten die Anzahl an Datensätzen in molekularbiologischen Datenbanken sehr stark an, und die Daten können von der Wissenschaftsgemeinschaft frei verwendet werden, etwa um Mustererkennung über mehrere, unabhängige Datensätze hinweg durchzuführen. Dabei spielen Data Science Methoden zum Fusionieren von Datensätzen eine wichtige Rolle.

Anfang des neuen Jahrtausends wurden Verfahren zur Hochdurchsatz-Sequenzierung von DNA- und RNA-Proben soweit entwickelt, dass die Sequenzierung einer einzigen biologischen Probe sehr kostengünstig wurde. Und wieder wuchsen die Datenbestände exponentiell weiter.

Mittlerweile sind „Systembiologen“ daran interessiert, die ganze Bandbreite molekularer Mechanismen und die Zusammenhänge zwischen Genom, Transkriptom, Proteom und anderen „Omics“-Ebenen genau zu verstehen. Dazu werden die oben beschriebenen großen Daten („Big Data“) häufig parallel an denselben Proben erhoben (Huang et al. 2017). Ziel einer Analyse sind dann nicht zwingenderweise einzelne Komponenten (wie etwa in der Genetik), sondern das gesamte Muster. Zur Auswertung werden viele der oben genannten Methoden der Data Science verwendet und von Bioinformatikern weiterentwickelt. Insofern sind die unterschiedlichen Disziplinen sehr aufeinander angewiesen und können stark voneinander profizieren.

Literatur

Huang S, Chaudhary K, Garmire LX (2017) More is better: recent progress in multi-omics data integration methods. Frontiers in genetics 8:84

Jung K, Vogel C, Zapf A, Frömke C (2019) Reproduzierbare Forschungsergebnisse: Anforderungen und Herausforderungen durch Data Science. mdi 21(2):45–48

Kruppa J, Jung K (2017) Automated multigroup outlier identification in molecular high-throughput data using bagplots and gemplots. BMC Bioinform 18:232

Was ist (Medical) Data Science? Eine integrative Perspektive

Prof. Dr. Antonia Zapf forscht als Biometrikerin zu statistischen Methoden für Diagnose- und Interventionsstudien, insbesondere zu adaptiven Designs. Neben der Forschung ist ihr das Selbstverständnis und die Fremdwahrnehmung der Biometrie ein wichtiges Anliegen.

Ich bin Biometrikerin (oder Biostatistikerin oder Medizinstatistikerin, aus meiner Sicht sind das Synonyme). Damit ist meine Aufgabe an einem Universitätsklinikum im Kontext der Wissenschaft die Weiterentwicklung der statistischen Methodik und die Anwendung von Statistik zur Analyse von Daten (Zapf et al. 2019). Wissenschaft mit Daten - also data science? Bin ich ein data scientist? Während der Begriff data science ursprünglich 1960 von dem dänischen Informatiker Peter Naur als bessere Alternative zum Begriff Informatik („computer science“) vorgeschlagen wurde, hat C.F. Jeff Wu 1997 data science mit Statistik gleichgesetzt (Ratner 2017). Schon anhand dieser Positionierungen wird klar, dass verschiedene Fachdisziplinen den Begriff data science für sich beanspruchen.

Vor kurzem habe ich mich auf einer Tagung mit Statistik-Studierenden unterhalten, die gerade am Übergang zum Beruf standen und mir von ihrer Stellensuche erzählten. Sie haben sich alle als data scientists bezeichnet und nach entsprechenden Ausschreibungen gesucht, waren allerdings ob der unübersichtlichen Situation frustriert. Kein Wunder: Wenn man sich entsprechende Stellenausschreibungen anschaut, findet man einen bunten Strauß von Aufgaben – von der Methodenentwicklung bis zur Visualisierung von Ergebnissen, von der Planung von Studien bis zur Modellierung von Daten, das Ganze gerne garniert mit den Begriffen big data und machine learning. Auch inhaltlich ist das Feld sehr heterogen: Stellenangebote für data scientists, umfassen Ausschreibungen von Banken, Wirtschaftsprüfungsgesellschaften, Beratungsfirmen und ähnlichem – die medizinische Forschung spielt hier eine verhältnismäßig kleine Rolle. Die Konsequenz ist eine allgemeine Konfusion: Wenn nur der Begriff data science ohne weitere Spezifikation auftaucht, dann wissen Bewerber nicht, was von ihnen erwartet wird und Arbeitgeber wissen bei dem heterogenen Bewerberfeld ebensowenig, was sie von data scientists erwarten können.

Im Kontext des vorliegenden Bandes ist aber genau das zu klären – was ist eigentlich medizinische Datenwissenschaft, d. h. medical data science? Um das zu klären, ist ein Blick auf die beteiligten Fachdisziplinen hilfreich: In diesem Gebiet der datengestützten Wissenschaft sind Bioinformatiker, Biometriker, Datenmanager, Epidemiologen, Medizininformatiker und noch weitere Berufsgruppen tätig. Erst aus der Zusammenschau der Profile dieser verschiedenen Disziplinen lässt sich ersehen, was medical data science alles umfasst.

Auf der einen Seite hat jede Disziplin ihre eigenen Kernkompetenzen und Verantwortlichkeiten – diese sollten die Kooperationspartner auch kennen, um Missverständnisse zu vermeiden und eine erfolgreiche Zusammenarbeit zu ermöglichen (Zapf et al. 2020). Daher halte ich es für richtig und hilfreich, die entsprechenden Berufsbezeichnungen zu verwenden. Auf der anderen Seite gibt es viele Überlappungsbereiche zwischen den Disziplinen. Daher halte ich es für absolut erstrebenswert, dass sich die verschiedenen Fachdisziplinen aus dem Bereich medical data science vernetzen, um die Synergien zu nutzen. Wenn wir es schaffen würden, ein stimmiges Gesamtkonzept zu entwickeln, bei dem jeder seine Expertise bestmöglich einsetzt und mit der der anderen Disziplinen verzahnt, würden wir ein leistungsstarkes und effizientes Konstrukt erhalten – medical data science mit den verschiedenen Berufsgruppen als Protagonisten. Die Realität sieht

allerdings im Moment noch so aus, dass häufig nicht nur um den Begriff data science, sondern auch um Studierende, Aufgaben, Stellen und Kooperationspartner gekämpft wird.

Dieses Buch ist ein schönes Beispiel für eine integrative Perspektive auf medical data science, die durch gleichzeitige Abgrenzung und Vernetzung der Disziplinen zustande kommt: Es gibt solche Beiträge, die zeigen, wie fachspezifische Lehrinhalte didaktisch gut vermittelt werden können und solche, die fachunspezifische, übergreifende didaktische Ansätze vorstellen. Die Verzahnung von beidem führt zu einer integrativen Perspektive auf Lehre im Kontext von medical data science.

In der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS), die über die namensgebenden Disziplinen hinaus auch die Disziplinen Medizinische Dokumentation, Bioinformatik und Systembiologie vertritt, steckt das Potential für einen Schulterschluss der Disziplinen, der einen integrativen Blick auf medical data science ermöglichen würde. Für diesen Schulterschluss sind aus meiner Sicht drei Schritte nötig: 1. Die klare Definition und gegenseitige Anerkennung der Kernkompetenzen und Verantwortlichkeiten der einzelnen Disziplinen, 2. die Identifizierung von Überlappungsbereichen und Nutzbarmachung für eine effiziente Zusammenarbeit und 3. die Kommunikation dieses Konstrukts von medical data science nach außen, gerichtet u. a. an Studierende, Kooperationspartner, Universitätsleitungen und Drittmittelgeber.

Was also hat es mit dem Begriff (medical) data science auf sich? Aus meiner Sicht kann (medical) data science alles sein und alle zugehörigen Disziplinen unter einen Hut bringen – aber (medical) data science ist nichts und bringt nichts, wenn jeder versucht, sich den Hut alleine aufzusetzen.

Literatur

Ratner B (2017) Statistical and machine-learning data mining: techniques for better predictive modeling and analysis of big data, 3. Aufl. Taylor & Francis

Zapf A, Huebner M, Rauch G, Kieser M (2019) What makes a biostatistician? Stat Med 38(4):695–701

Zapf A, Rauch G, Kieser M (2020) Why do you need a biostatistician? BMC Med Res Methodol 20(1):23

Inhaltsverzeichnis

Statistischer Humor im Unterricht

1

Witze und Cartoons für signifikanten Spaß mit relevantem Effekt

Annette Aigner

1.1 Einleitung

F:„Statistik?“ A: „Das mochte ich noch nie.“
Viele Dozierende in sozialwissenschaftlichen Fächern, Psychologie, Epidemiologie, Public Health, und vielen mehr sind mit der großen Aufgabe konfrontiert, ein tendenziell a priori mit negativen Assoziationen besetztes, unbeliebtes, verstaubt und trocken geglaubtes, und für unverständlich und unzugänglich gehaltenes Pflichtfach zu unterrichten. Dies ist natürlich keine einfache Aufgabe und erfordert konstantes Engagement und Motivation der Dozierenden – potenziell über mehrere Unterrichtseinheiten und Semester hinweg.

Es gibt viele Belege dafür, dass Humor in der Lehre für die Qualität des Unterrichts und die Ergebnisse der Studierenden funktional relevant ist (Wanzer et al. 2010). Und mit Humor im Statistikunterricht hat man die Möglichkeit, Stereotypen oder Unbehagen zu zerstreuen – gegenüber dem Fach, dem Kurs oder gar den Dozierenden (Lesser und Pearl 2008).

Elektronisches Zusatzmaterial Die elektronische Version dieses Kapitels enthält Zusatzmaterial, das berechtigten Benutzern zur Verfügung steht https://doi.org/10.1007/978-3-662-62193-6_1.

A. Aigner (✉)
Institut für Biometrie und Klinische Epidemiologie, Universitätsmedizin Berlin – Charité, Berlin, Deutschland
E-Mail: annette.aigner@charite.de

C. Herrmann et al. (Hrsg.), *Zeig mir Health Data Science!*,
https://doi.org/10.1007/978-3-662-62193-6_1