



## Digitale Epidemiologie

### Einleitung

Vor etwa 20 Jahren wurde das erste Ganzgenom eines multizellulären Organismus, des Fadenwurms *Caenorhabditis elegans* veröffentlicht [1]. Nur sechs Jahre später wurde im Rahmen des *Human Genome Projects* ein substantieller Teil des menschlichen Genoms sequenziert [2]. Seitdem sind nicht nur die Kosten, sondern auch der technische und zeitliche Aufwand für die Sequenzierung um mehrere Größenordnungen gefallen [3]. Molekulargenetische Ansätze sind heutzutage in zahlreichen Bereichen außerhalb der Biologie zu einem Standardwerkzeug geworden, haben diverse wissenschaftliche Gebiete revolutioniert und einen enormen, in einigen Bereichen disruptiven Erkenntnisgewinn gebracht [4]. Neuere Methoden wie Next Generation Sequencing (NGS) erlauben es mittlerweile, die organismische Zusammensetzung einzelner Proben über ihren genetischen „Footprint“ zu messen (Metagenomik; [5]).

Parallel zur experimentellen Datengewinnung hat sich der Wissenschaftszweig Bioinformatik zum am stärksten wachsenden Teilgebiet der Informatik entwickelt. Da immer größere Datenmengen durch effizientere Sequenzierung gewonnen werden, müssen hochperformante Computeralgorithmen entwickelt werden, die aus den Rohdaten relevante Informationen extrahieren, Datenlücken mit hoher Verlässlichkeit schließen, effektives Management gewährleisten und große Datenmengen klassifizieren.

In der Epidemiologie ist die Verzahnung molekulargenetischer und bioinformatischer Methoden nicht mehr weg-

zudenken. In verschiedensten Anwendungen werden traditionelle Ansätze, zum Beispiel in der Infektionsepidemiologie übertragbarer Krankheiten, durch genetische Informationen über Erreger ergänzt, erweitert oder verbessert. So können die schnelle Sequenzierung von Proben und ihre bioinformatische Analyse im Kontext nosokomialer Ausbrüche antibiotikaresistenter Keime Aufschluss über potenzielle Transmissionswege innerhalb eines Krankenhauses liefern [6].

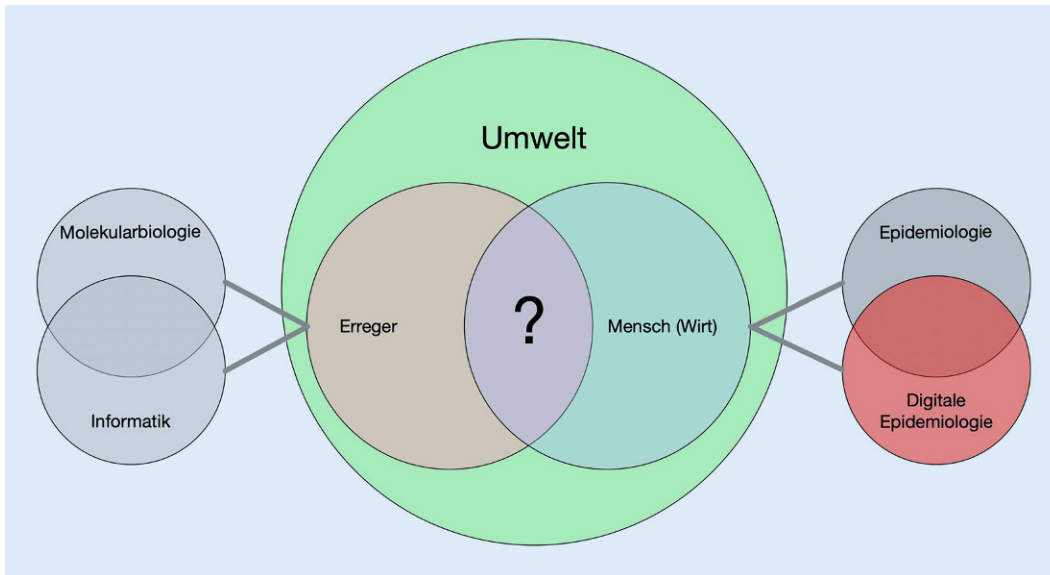
Mittlerweile ist die transdisziplinäre Brücke zwischen Molekularbiologie und Bioinformatik eine solide wissenschaftliche Verbindung in der Epidemiologie und ein gutes Beispiel für Synergien, die durch die Vernetzung vormals disjunkter Wissenschaftsgebiete entstehen. Oftmals werden diese Vernetzungen durch die Entwicklung neuer Technologien geradezu erzwungen. Man denke nur an die Entwicklung des Mikroskops vor 400 Jahren, das gewissermaßen die Medizin mit der Biologie verschränkt und den Lebenswissenschaften eine unvergleichliche Kaskade von Erkenntnissen gebracht hat.

### Digitale Epidemiologie

Momentan werden wir wieder Zeugen einer vergleichbaren wissenschaftlichen Revolution, die durch neue Technologien ausgelöst wurde und immer noch anhält. Diese Revolution wird im Wesentlichen durch drei technologische Errungenschaften der letzten Jahre getragen und vorangetrieben: 1.) das Internet, 2.) soziale Netzwerke und 3.) mobile Endgeräte. Diese Technologien sind das technologische Fundament der digitalen Epidemiologie.

Der Begriff „Digital Epidemiology“ wurde 2012 von Marcel Salathé geprägt, um Forschungsprojekte und -aktivitäten im Bereich der Epidemiologie zusammenzufassen, bei denen eine oder mehrere dieser Schlüsseltechnologien nutzbar gemacht werden [7]. Ähnlich wie die Schlagworte „Big Data“, „Data Analytics“, „maschinelles Lernen“, „künstliche Intelligenz“, „Digital Health“ etc. führt der Begriff „Digital Epidemiology“, zu Deutsch „digitale Epidemiologie“, oftmals zu leichter Verwirrung, wird nicht selten missverstanden oder mit anderen Konzepten verwechselt. Hinzu kommt, dass Schlagworte auch immer vage bleiben und die Inhalte, die sich dahinter verbergen, gerade hierzulande immer auch etwas bedrohlich wirken können, weil sie unbekannt sind. Dieser Effekt wird noch dadurch verstärkt, dass gerade einige Konzepte, wie z. B. „maschinelles Lernen“ oder „künstliche Intelligenz“, in Teilen der Gesellschaft negativ besetzt sind, in der digitalen Epidemiologie aber eine wichtige Rolle spielen.

Wieso das noch neue, rasant wachsende Gebiet „digitale Epidemiologie“ so vielversprechend ist, kann gut am Beispiel von Ausbrüchen direkt übertragbarer Infektionskrankheiten erläutert werden. Jedes einzelne Ausbruchsgeschehen ist ein komplexes dynamisches Phänomen, bei dem eine Vielzahl von Faktoren eine Rolle spielt. Trotz spezifischer Unterschiede eint alle Prozesse die Tatsache, dass sie an der Schnittstelle zwischen Erreger und Wirt stattfinden und dass Erreger und Wirt systemspezifischen Umweltbedingungen ausgesetzt sind (siehe **Abb. 1**). Biologische Faktoren des Erregers spielen eine wichtige Rolle und werden z. B. durch molekulargenetische Me-



**Abb. 1** ◀ Methoden zur Erforschung von Wechselwirkungen zwischen Erreger und Wirt in der Umwelt bei direkt übertragbaren Infektionskrankheiten. Die digitale Epidemiologie bietet neue Technologien zum Verständnis der wirtsseitigen Faktoren

Methoden in Verbindung mit bioinformatischen Analysen bestimmt. Das Schlüsselproblem liegt auf der Seite des Wirts, also des Menschen. Die Vielzahl der Faktoren, die wirtsseitig eine Rolle spielen, lassen sich praktisch nicht quantitativ erfassen. Eine ganze Reihe individueller Verhaltensmuster bestimmt Transmissionswahrscheinlichkeiten, die mechanistisch nicht einmal ausreichend beschrieben werden können. Insbesondere bei übertragbaren Krankheiten ist es aber notwendig, die Struktur der Interaktionen zwischen Individuen zu messen, die wiederum stark kontextabhängig sind. Mobilitätsmuster können eine Rolle spielen und bestimmen, wie stark eine Population durchmischt wird. Modelle zeigen, dass die Strukturen von Kontaktnetzwerken den Verlauf einer Epidemie stark beeinflussen können [8]. Diese Netzwerke sind allerdings nur schwer zu erfassen, insbesondere in natürlichen Verhaltenssituationen. Es können eben keine systematischen Experimente durchgeführt werden. In den meisten Fällen ist man auf Surveys und indirekte Methoden angewiesen.

Die digitale Epidemiologie setzt an dieser Stelle an. Mithilfe der o.g. Technologien ist es teilweise möglich, wichtige Erkenntnisse über das menschliche Verhalten zu gewinnen, besonders über Entscheidungsprozesse, Bewegungsmuster und, am allerwichtigsten, die Wechselwirkungen zwischen Individuen, die

mit konventionellen Methoden nicht erfasst werden können. Diese Forschung steht noch am Anfang, aber schon jetzt zeichnet sich ab, dass die digitale Epidemiologie für das Verständnis der Dynamik von Infektionskrankheiten eine ebenso wichtige Rolle spielen wird wie die Bioinformatik für das Verständnis des Erregers.

Dieser Artikel soll einen Überblick liefern. Fragen wie: „Welche Methoden kommen in der digitalen Epidemiologie zum Einsatz?“, „Welche Art von Daten werden analysiert?“ und „Was sind die typischen Datenquellen?“ werden beantwortet und an Beispielen diskutiert. Ein besonderer Fokus wird auf die Chancen dieses neuen Forschungszweiges gelegt. Es werden die Herausforderungen diskutiert und erläutert, wie auch in diesem Bereich der Erfolg im Kern davon abhängt, wie mutig Wissenschaftler mit bisher unbekanntem Methoden, Ansätzen und Sichtweisen umgehen können.

## Neue Datenquellen

### Datenintegration neuer Qualität

Ein wichtiger Strang der digitalen Epidemiologie ist die Erweiterung klassischer Surveillance-Systeme. Insbesondere die Analyse von Inzidenz- und Prävalenzdaten wird immer stärker durch digitale Technologien verbessert und verfeinert. Hocho aufgelöste Daten mit breiter Ab-

deckung werden mittlerweile integrativ auf öffentlich zugänglichen Datenportalen zur Verfügung gestellt. Die Internetplattform Healthmap (<https://www.healthmap.org/en>) ist hierfür ein gutes Beispiel [9]. Ziel dieser Plattform ist, verschiedenste Datenquellen (WHO [World Health Organization], ProMed Mail [10], GeoSentinel [11], OIE [12], FAO [13], EuroSurveillance [14], etc.) zu Ausbrüchen verschiedener Erreger in einer Datenbank zusammenzufassen. Sowohl Infektionskrankheiten als auch Zoonosen und andere Krankheiten sind integriert. Daten werden auf dieser Plattform nicht nur gesammelt und zur Verfügung gestellt, sondern auch interaktiv visualisiert (siehe **Abb. 2**). Healthmap ist ein hervorragendes Beispiel für die Zusammenarbeit von Epidemiologen, Medizinern, Biologen, Sozialwissenschaftlern und Informatikern. Die Integration verschiedener Datenquellen in einer Gesamtdatenbank birgt den offensichtlichen Vorteil, viel schneller Ausbruchsgeschehen automatisiert zu identifizieren. So können speziell dafür konzipierte Algorithmen statistische Unregelmäßigkeiten detektieren und von spontanen Schwankungen unterscheiden. Mittlerweile ist Healthmap zu einem Werkzeug gereift, das in verschiedenen wissenschaftlichen Studien als zentrale Datenquelle herangezogen wird, so z. B. in einer vielzitierten Untersuchung zur globalen Verteilung von Dengue [15].

Ein zweites erfolgreiches Beispiel ist das Projekt NextStrain (<https://nextstrain.org/>), konzipiert und entwickelt von Trevor Bedford und Richard Neher [16]. NextStrain ist eine Internetplattform, die molekularbiologische Informationen zur Evolution verschiedener Virusinfektionskrankheiten (Influenza, Masern, Ebola, West-Nil-Virus etc.), aber mittlerweile auch bakterieller Infektionskrankheiten mit geografischen Informationen vereint. NextStrain ist damit im neuen Gebiet der Phylogeografie angesiedelt (■ **Abb. 3**). Genau wie Healthmap integriert NextStrain Daten aus verschiedensten Quellen. Das Tool wird mittlerweile von Wissenschaftlern genutzt, um z. B. die weitere Evolutionsdynamik verschiedener Viren vorherzusagen oder um die Zusammensetzung von Impfstoffen zu optimieren [17]. Wie bei vielen Projekten dieser Art wird auch hier Wert darauf gelegt, alle zum Einsatz kommenden Softwarekomponenten unter Public License und Open Source auf Code-Sharing-Plattformen zur Verfügung zu stellen.

### Soziale Netzwerke als Datenquelle

Epidemiologie ist *per definitionem* eine interdisziplinäre Wissenschaft, in der Expertenwissen aus Medizin, Biologie, Sozialwissenschaften und Statistik zusammengeführt wird. Klassisch werden Daten aus der Surveillance mit statistischen Methoden untersucht, statistische Modelle werden entwickelt, um Inzidenz und Prävalenz einer Infektionskrankheit vorherzusagen. Sowohl die Qualität der Surveillance als auch die Genauigkeit der zum Einsatz kommenden statistischen Methoden hat sich graduell immer weiter verbessert.

Eine vielversprechende, ergänzende Komponente hat sich durch die Integration sozialer Medien und sozialer Netzwerke ergeben. Wissenschaftler aus verschiedenen Disziplinen entwickeln mittlerweile Methoden, um aus individuen-spezifischen Daten, die gewissermaßen als Nebenprodukt auf Internetplattformen sozialer Netzwerke gewonnen werden, Antworten auf epidemiologische Fragestellungen zu finden. Netzwerke wie Facebook oder Twitter

Bundesgesundheitsbl 2020 · 63:166–175 <https://doi.org/10.1007/s00103-019-03080-z>  
© Der/die Autor(en) 2020

D. Brockmann

## Digitale Epidemiologie

### Zusammenfassung

Digitale Epidemiologie ist ein relativ neues, rapide wachsendes Forschungsgebiet. Die technologische Revolution des letzten Jahrzehnts, die globale Vernetzung, der Informationsaustausch über soziale Medien und insbesondere die nahezu gesamtgesellschaftliche Nutzung mobiler Endgeräte ermöglichen es, individuen-spezifische Daten mit einer Auflösung zu erfassen, die noch vor einigen Jahren unmöglich erschien. Diese Daten werden in der *digitalen Epidemiologie* untersucht, um epidemiologische Fragen besser beantworten zu können. Dieser Artikel liefert einen Überblick. Es werden verschiedene Aspekte der digitalen Epidemiologie diskutiert. An Beispielen wird erläutert, wie epidemiologische und bioinformatische Daten auf interaktiven Internetplattformen zusammengeführt werden, wie durch Analyse der Inhalte und des Informationsaustauschs über soziale Medien und Netzwerke wichtige Erkenntnisse gewonnen werden und wie mithilfe mobiler Endgeräte in natürlichen Experimenten Kon-

takt- und Proximitätsnetzwerke rekonstruiert werden, um die Dynamik direkt übertragbarer Infektionskrankheiten besser verstehen, beschreiben und vorhersagen zu können. Es wird erklärt, wieso die moderne Netzwerktheorie, aber auch Methoden des maschinellen Lernens und künstliche Intelligenz bei der Analyse sehr großer Datensätze wichtige Werkzeuge sind und wie traditionelle, statistische Ansätze der Infektionsepidemiologie durch diese neuen Methoden ergänzt werden. Die ethischen Herausforderungen im Bereich Datenschutz, Datensicherheit und Persönlichkeitsrechte werden schließlich diskutiert. Konzepte und Wege, personenbezogene Verhaltensdaten einerseits nutzbar zu machen und andererseits die Datenhoheit jedes Einzelnen zu wahren, werden skizziert.

### Schlüsselwörter

Komplexe Netzwerke · Maschinelles Lernen · Künstliche Intelligenz · Big Data · Computational Epidemiology

## Digital epidemiology

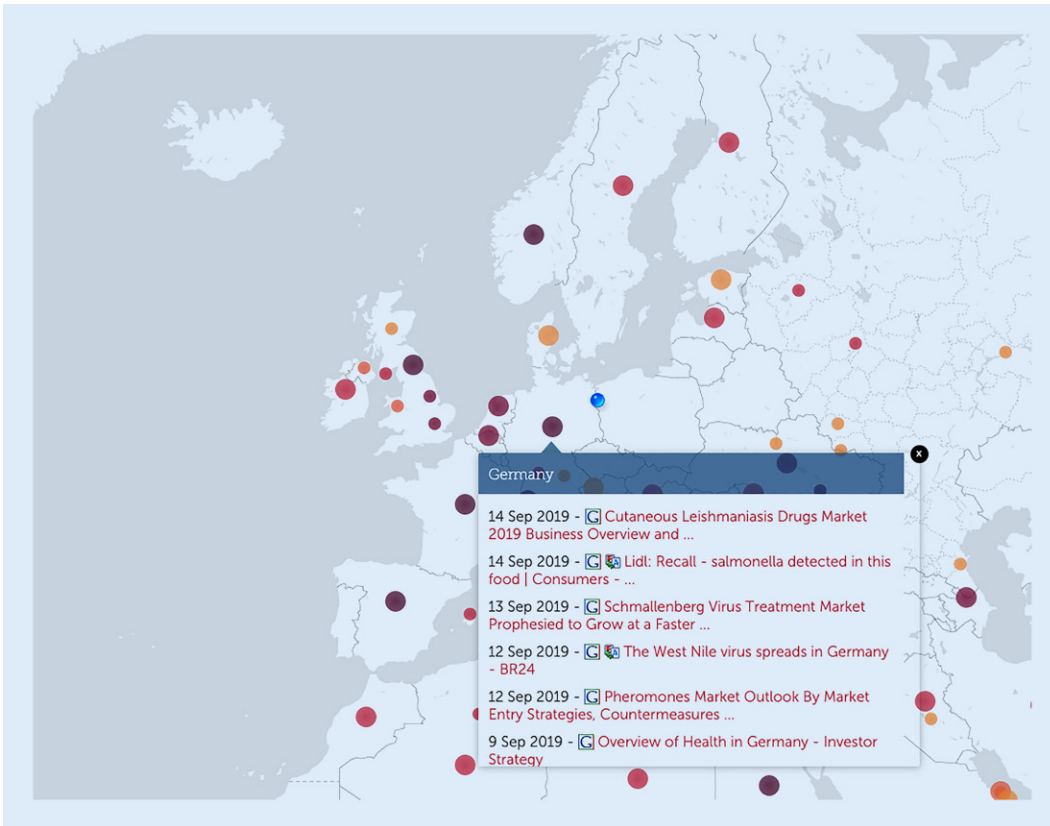
### Abstract

Digital epidemiology is a new and rapidly growing field. The technological revolution we have been witnessing during the last decade, the global rise of the Internet, the emergence of social media and social networks that connect individuals worldwide for information exchange and social interactions, and the almost complete social penetration of mobile devices such as smartphones provide access to data on individual behavior with unprecedented resolution and precision. In digital epidemiology, this type of high-resolution behavioral data is analyzed to advance our understanding of, for example, infectious disease dynamics and improve our abilities to forecast epidemic outbreaks and related phenomena. This article provides an overview on the topic. Different aspects of digital epidemiology are alluded to. Based on examples, I will explain how epidemiological data is integrated on new comprehensive and interactive websites,

how the analysis of interactions and activities on social media platforms can yield answers to epidemiological questions, and finally how individual-based data collected by smartphones or wearable sensors in natural experiments can be used to reconstruct contact and physical proximity networks the knowledge of which substantially improves the predictive power of computational models for transmissible infectious diseases. The challenges posed in terms of privacy protection and data security will be discussed. Concepts and solutions will be explained that may help to improve public health by leveraging the new data while at the same time protecting the individual's data sovereignty and personal dignity.

### Keywords

Complex networks · Machine learning · Artificial intelligence · Big data · Computational epidemiology



**Abb. 2** ◀ Interaktive Visualisierung von Daten zu verschiedenen Infektionskrankheiten im öffentlich zugänglichen Datenportal Healthmap ([www.healthmap.org](http://www.healthmap.org))

aggregieren Daten, die Rückschlüsse über die Dynamik von Infektionskrankheiten erlauben oder mit überraschender Genauigkeit z. B. den Verlauf einer saisonalen Grippewelle reproduzieren können [18]. So können z. B. der Effekt von Impfkampagnen und die Impfbereitschaft in Populationen quantitativ über die Textanalyse der Inhalte auf sozialen Medienplattformen erfasst werden [19]. Obwohl diese Entwicklung noch nicht ausgereift ist, sind erste Ergebnisse sehr vielversprechend und werden sich sicher zu einer wichtigen Ergänzung traditioneller und klassischer Surveillance-Systeme entwickeln. Gerade die Daten des Kurznachrichtendienstes Twitter werden immer interessanter für Wissenschaftler aus dem Bereich Public Health. Zum Beispiel konnte in einer Reihe jüngst veröffentlichter Studien die Vernetzung sowohl von Impfgegnern als auch Impfbefürwortern in den USA quantitativ erfasst und ausgewertet werden. Außerdem wurde sehr präzise ausgewertet, welche Onlinemedien die jeweiligen Gruppen nutzen, um Informationen auszutauschen und informiert

zu werden [20]. Diese Daten liefern dann weitere Erkenntnisse, wenn moderne netzwerktheoretische Methoden zum Einsatz kommen und die Existenz sogenannter Echo-Chambers (Gruppen von Gleichgesinnten; [21]) nachweisen können, um die Persistenz nichtevidenzbasierter Meinungen in sozialen Gruppen zu erklären oder geografisch zu lokalisieren. Hierbei kann gut erkannt werden, an welchen Stellen z. B. Informationskampagnen greifen könnten.

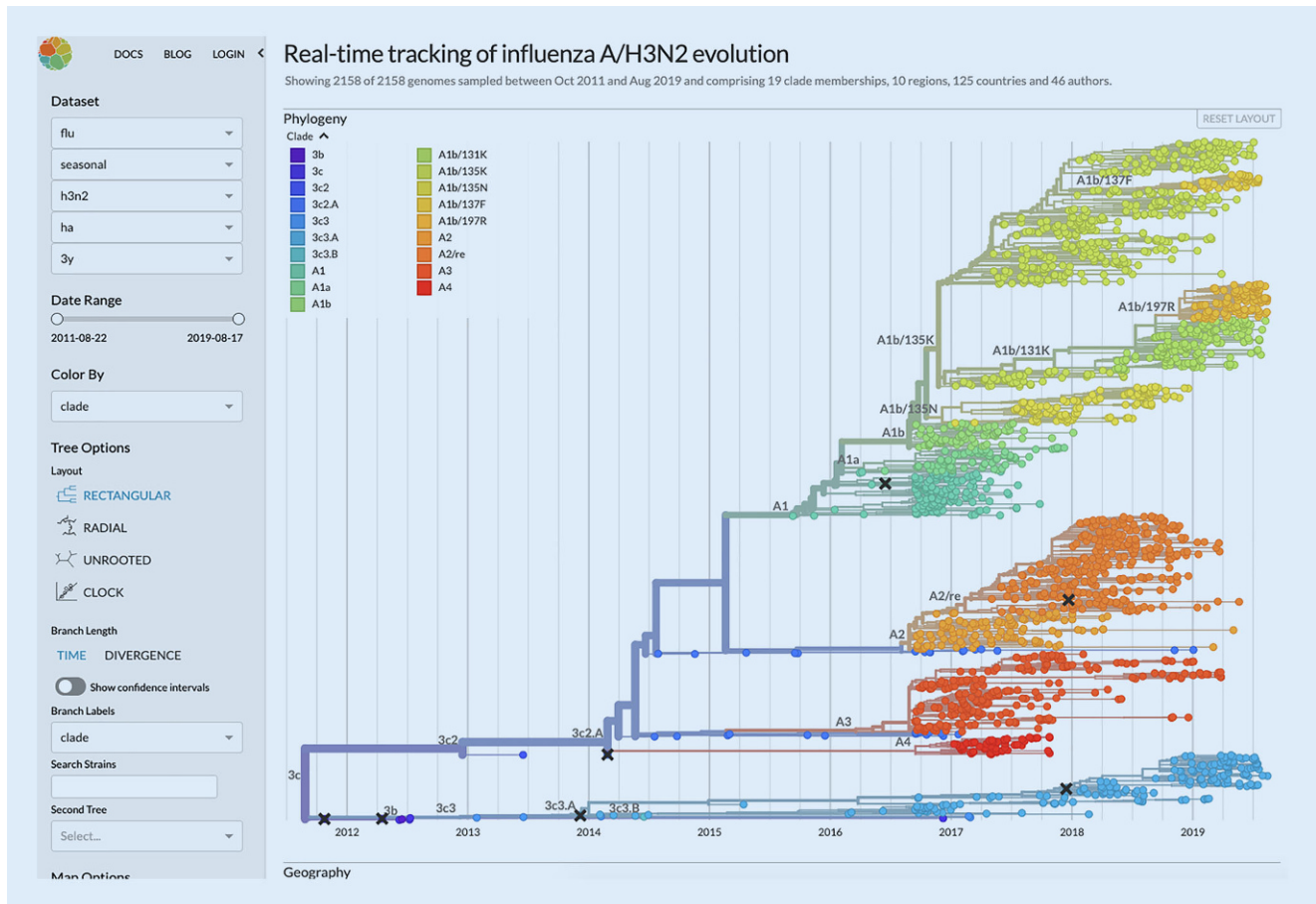
### Mobile Endgeräte, Smartphones und tragbare Sensoren

Eine der fundamentalen Schwierigkeiten in der Infektionsepidemiologie übertragbarer Krankheiten ist die Rekonstruktion individueller Transmissionsgeschehen bzw. Transmissionswege in einer Population. Aus reinen Inzidenzen bzw. Prävalenzen können diese nicht rekonstruiert werden, was die Vorhersage des weiteren Verlaufs erschwert oder unmöglich macht. Phylogenetische Analysen können helfen, erfordern aber Zeit und können nur bedingt eingesetzt

werden. Im Kern liegt das Problem darin, dass die Dynamik direkt übertragbarer Krankheiten durch die zugrunde liegenden, direkten Wechselwirkungen von Individuen bestimmt wird, deren genaue Messung über längere Zeit und in repräsentativen Populationen unter natürlichen Bedingungen aber unmöglich ist.

In einem Gedankenexperiment könnte man jeden physischen Kontakt oder jede räumliche Proximität von Personen messen, um in Verbindung mit dem Infektionszustand jeder Person potenzielle Infektionen entweder zu identifizieren oder auszuschließen. Gerade am Beispiel direkt übertragbarer Infektionskrankheiten wird klar, dass statt der zeitlichen Inzidenz- oder Prävalenzdynamik der zeitliche Verlauf der Interaktionen zwischen infektiösen und suszeptiblen Personen analysiert werden müsste, was in der Sprache der Netzwerktheorie bedeutet: „Es kommt nicht auf die Netzwerknoten an, sondern auf deren Verbindungen.“

An diesem Punkt setzt ein weiterer Teilbereich der digitalen Epidemiologie

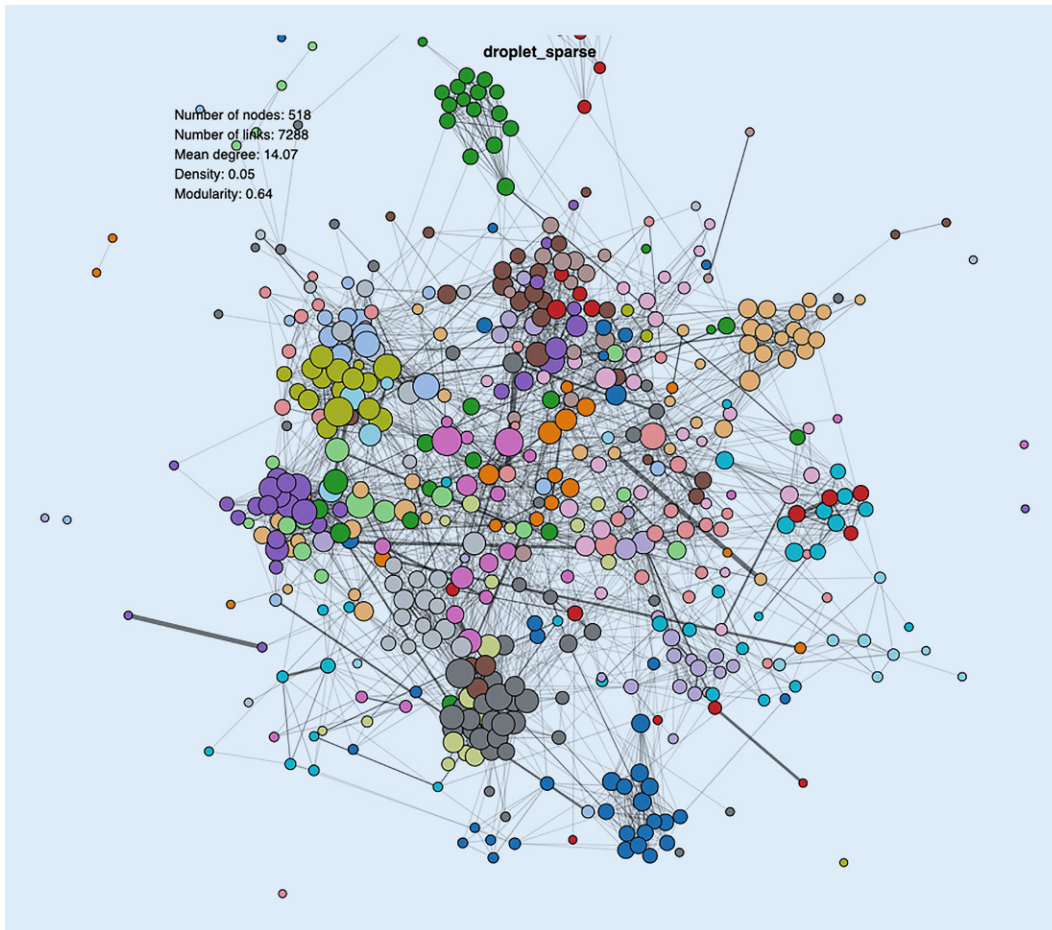


**Abb. 3** ▲ Visualisierung von genetischen und geografischen Daten zu verschiedenen Erregern im öffentlichen Internetportal NextStrain ([www.nextstrain.org](http://www.nextstrain.org))

an: die genaue Erfassung einzelner Interaktionen zwischen Individuen in Populationen mittels neuer individuenbasierter Sensoren. Schon 2008 haben Wissenschaftler um Alain Barrat, Ciro Cattuto und Alessandro Vespignani Experimente durchgeführt, bei denen Probanden in ausgewählten, aber natürlichen Szenarien mit tragbaren RFID („radio-frequency identification“-)Chips ausgestattet wurden [22]. Das SocioPatterns-Projekt war geboren ([www.sociopatterns.org](http://www.sociopatterns.org)). Die tragbaren Sensoren konnten mit hoher Auflösung die physische Proximität zwischen Individuen messen. Immer, wenn sich zwei Personen in einem Abstand von weniger als einem Meter gegenüberstanden, wurde ein Kontakt gemessen. Die Wissenschaftler hatten zunächst das Ziel, die statistischen, temporalen und topologischen Strukturen der Interaktionsnetzwerke zu verstehen. Experimente wurden in Museen, Schulen und auf Konferenzen durchgeführt [23–25].

Das Potenzial dieser Methode zur Kontaktnetzwerkrekonstruktion im Public-Health-Bereich wurde schnell erkannt und verschiedene Wissenschaftler haben diese Technologie im Kontext nosokomialer Infektionen eingesetzt [26]. In einer Studie von 2013 wurden in einem Krankenhaus Patienten und Krankenhausangestellte mit RFID-Chips ausgestattet. Mit einer zeitlichen Auflösung im Sekundenbereich wurde über mehrere Wochen jeder Kontakt zwischen einigen Hundert Probanden gemessen [27]. Mithilfe dieser Daten konnte gezeigt werden, wie sich Kontakthäufigkeiten zwischen den verschiedenen Gruppen von Personen (z. B. zwischen Patienten und Personal) quantitativ unterscheiden und von welchen Parametern Unterschiede abhängen. Mithilfe netzwerktheoretischer Methoden konnten die aggregierten Kontaktnetzwerke auf strukturelle Muster und statistisch auffällige Eigenschaften untersucht werden.

In einer ähnlichen Serie von Studien hat eine Forschungsgruppe an der DTU (Technische Universität Kopenhagen) unter Leitung von Sune Lehmann im Rahmen des SensibleDTU-Projekts im Jahr 2012 ca. 1000 Smartphones an Studenten in einem partizipatorischen Experiment verteilt [28]. Die einzelnen Geräte wurden mit einer speziellen Software ausgestattet, die alle Aktivitäten jedes Individuums aufzeichnet und in einer Datenbank sammelt. Sowohl Aktivitäten auf sozialen Medien, der Austausch von SMS und der genaue Aufenthalt via GPS wurden erfasst und über viele Monate mit einer zeitlichen Auflösung von einigen Minuten gemessen. Insbesondere wurde via Bluetooth erfasst, wann sich zwei Personen der Kohorte in physischer Proximität befanden und wie lange diese Kontakte andauerten. **Abb. 4** zeigt ein über drei Monate aggregiertes Kontaktnetzwerk, das aus diesen Experimenten gewonnen wurde. Jeder Netzwerkknoten



**Abb. 4** ◀ Struktur eines Kontaktnetzwerkes, das aus SensibleDTU-Daten rekonstruiert wurde. Jeder Knoten repräsentiert eine Person. Die Verbindungen quantifizieren die Häufigkeit, mit der zwei Personen über einen Zeitraum von drei Monaten in Kontakt waren. Die Farben unterscheiden berechnete Cluster

repräsentiert eine Person. Die Netzwerkverbindungen, die sog. Links, quantifizieren, wie häufig und lange zwei Personen in näherem Kontakt waren. Man erkennt sofort, dass sich diese Populationen nicht durchmischen. Es existieren Cluster mit starker interner Vernetzung, die untereinander nur schwach verknüpft sind. Die Intensität der Verbindungen variiert ebenfalls stark. In verschiedenen Folgestudien und Modellen wurde gezeigt, welchen substanziellen Einfluss diese versteckten Strukturen realer Kontaktnetzwerke auf die Ausbreitung von Infektionskrankheiten haben können [29]. Es konnte in diesen Experimenten zweifelsfrei festgestellt werden, dass Modelle, die z. B. eine homogene Durchmischung annehmen oder die zeitlichen Modulationen in Netzwerkstrukturen ignorieren, zu stark fehlerhaften Vorhersagen kommen [30, 31]. Diese Einsicht ist insbesondere deshalb wichtig, weil konventionelle Modelle, mithilfe derer Größen wie die notwendige Durchimpfung zur

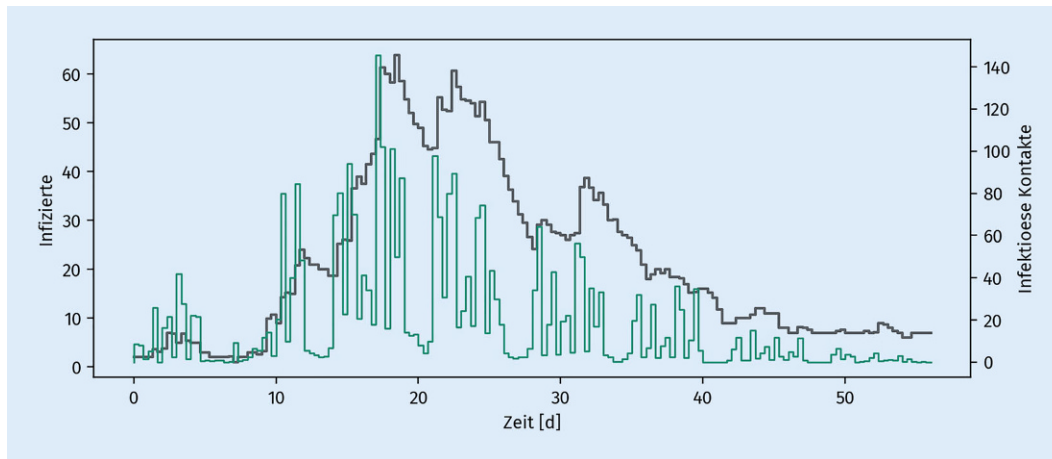
Auslöschung von Erkrankungen, wie z. B. Masern, berechnet werden, teilweise auf Annahmen fußen, die nicht einmal approximative Gültigkeit haben und deshalb neu evaluiert werden müssen.

Wie wichtig die Messung von Interaktionen und damit potenziellen Transmissionen ist, zeigt **Abb. 5**. Die Abbildung illustriert den Verlauf einer simulierten Infektionskrankheit auf der Datengrundlage des SensibleDTU-Experiments. Das heißt, hier wurde eine hypothetische Infektionskrankheit simuliert, bei der Transmissionen durch die empirisch ermittelten Kontakte stattfinden. Während die konventionelle Inzidenzkurve einen charakteristischen Verlauf mit exponentiellem Anstieg und darauffolgendem Abfall hat, ist der zeitliche Verlauf der Kontakte zwischen infizierten und suszeptiblen Personen extrem starken Schwankungen unterworfen. Es ist aber genau diese Kurve, die das Ergebnis bestimmt. Die starken Fluktuationen weisen auch darauf hin,

dass gemittelte Größen, wie die traditionell verwendete Basisreproduktionszahl (die mittlere Anzahl von Sekundärinfektionen, die eine infizierte Person für die Dauer der Infektion in einer vollständig empfänglichen (suszeptiblen) Population im Durchschnitt auslöst), nur mit großer Vorsicht und Skepsis für quantitative Vorhersagen herangezogen werden sollten.

## Methoden in der digitalen Epidemiologie

In der digitalen Epidemiologie kommen verschiedene Methoden zum Einsatz, die sich besonders zur Analyse der neu gewonnenen Daten eignen. Hier sollen drei Methoden schwerpunktmäßig näher diskutiert werden, die in der Infektionsepidemiologie noch nicht zu den Standardwerkzeugen gehören, aber gerade im Bereich direkt übertragbarer Krankheiten traditionelle Methoden sehr gut ergänzen können.



**Abb. 5** ◀ Verlauf einer hypothetischen Infektionskrankheit, die durch direkte Kontakte übertragen wird, simuliert auf Basis der empirischen Kontaktnetzwerke aus den SensibleDTU-Messungen. *Grau*: traditionelle Inzidenzkurve: Anzahl der Neuinfektionen pro Zeiteinheit. *Grün*: Verlauf der Anzahl potenziell infektiöser Kontakte

## Komplexe Netzwerke

Die o. g. Beispiele zeigen, dass die moderne Netzwerkforschung ein ganz wichtiges Werkzeug der digitalen Epidemiologie ist. Die „natürlichen Experimente“, exemplarisch repräsentiert durch die Projekte SensibleDTU und SocioPatterns, messen mit hoher Präzision Kontaktnetzwerke zwischen Individuen in Populationen unter natürlichen Bedingungen. Um diese oftmals zeitabhängigen Netzwerke besser verstehen und wichtige Strukturen extrahieren zu können, liefern netzwerktheoretische Methoden wichtige Ergebnisse. So kann z. B. die Variabilität des Knotengrads (der Knotengrad ist die Anzahl der Verbindungen eines Netzwerk-knotens) als ein Maß für die Heterogenität des Netzwerks Aufschluss geben, inwieweit sich epidemiologische Größen im Vergleich zu einer gut durchmischten Population verschieben [8, 32]. Typischerweise sind soziale Kontaktnetzwerke stark „geclustert“, d. h., es existieren stark vernetzte Bereiche des Netzwerks, die untereinander nur schwach verlinkt sind. Starkes Clustering kann dazu führen, dass Infektionskrankheiten sich zwar langsamer ausbreiten, aber auch persistenter sind [33].

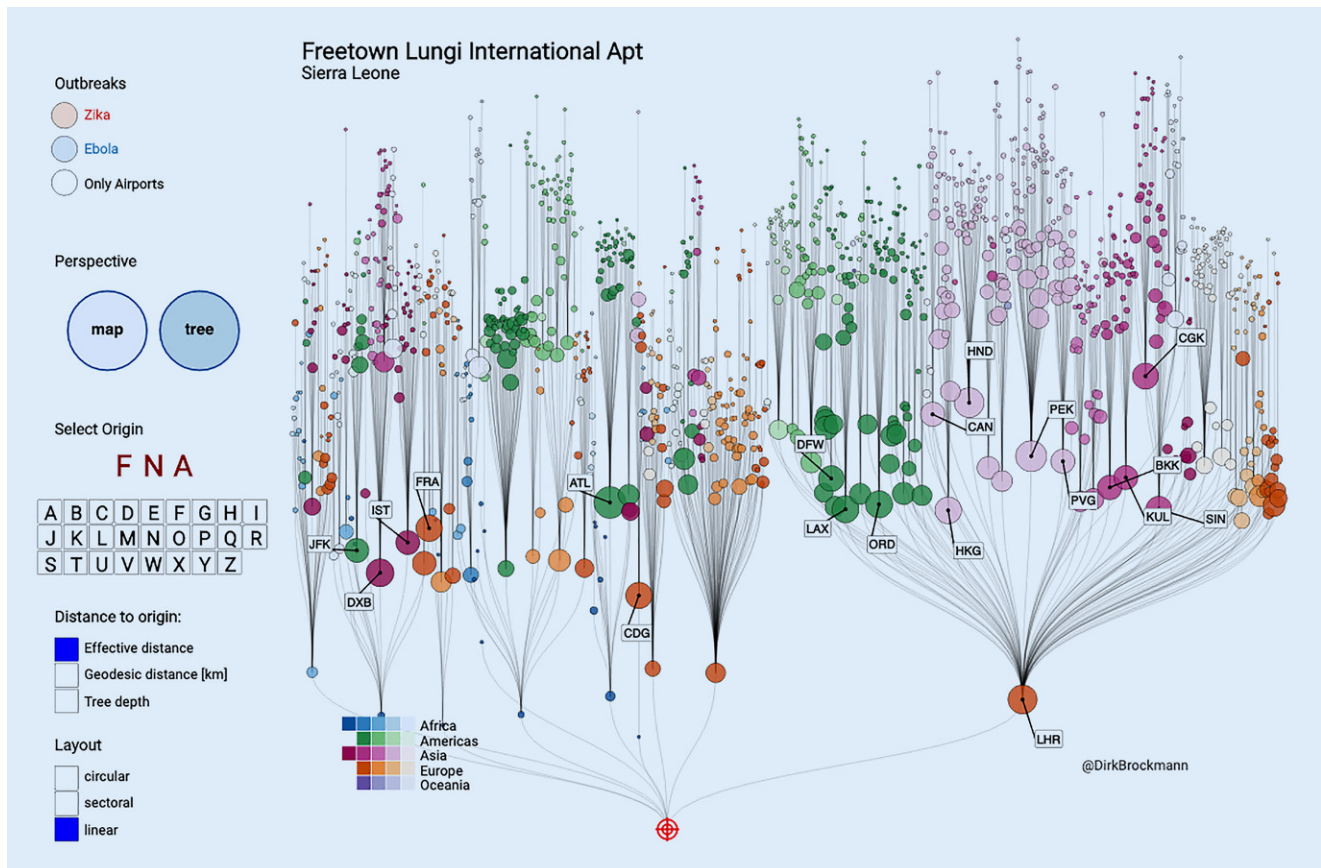
Die Netzwerktheorie ist ebenso wichtig für das Verständnis der geografischen Ausbreitung von Infektionskrankheiten, bei der natürlicherweise Mobilität eine Rolle spielt. Während noch vor einigen Jahren quantitative Mobilitätsmessungen für große Populationen unmöglich waren, liefern jetzt die neuen Technologien, insbesondere mobile Endgeräte

und interaktive Kartenanwendungen wie Google-Maps (Google LLC, Mountain View, CA, USA), Open-Streetmaps etc., sehr wertvolle Daten zu Bewegungsmustern einzelner Personen und gesamter Populationen. Aus diesen Bewegungsmustern können Netzwerke rekonstruiert werden, bei denen einzelne Knoten Orte repräsentieren und deren Verlinkung die Anzahl von Personen, die sich pro Zeiteinheit zwischen den Orten bewegen. Diese Mobilitätsnetzwerke bestimmen die geografische Ausbreitung von Infektionskrankheiten. Ihre Kenntnis verbessert Vorhersagen substantziell. **Abb. 6** zeigt das weltweite Flugverkehrsnetz und die Ergebnisse eines Netzwerkmodells zur Ausbreitung von Ebola im Jahr 2013. Das Modell wurde entwickelt, um die wahrscheinlichsten Ausbreitungswege zu bestimmen und das relative Importrisiko an verschiedenen Flughäfen weltweit zu schätzen. Durch mobile Endgeräte und deren GPS-Sensoren können mittlerweile individuelle Bewegungsmuster mit extrem hoher zeitlicher und räumlicher Auflösung gewonnen werden. Diese Bewegungsdaten können auch genutzt werden, um die Dynamik von Infektionskrankheiten in urbanen Settings besser verstehen und modellieren zu können.

## Interaktive Visualisierung

Neben der Netzwerktheorie werden auch immer stärker interaktive Visualisierungen als Werkzeug eingesetzt. Die interaktiven Komponenten z. B. in den Datenportalen Healthmap und NextStrain sind

ein wesentliches Merkmal dieser Tools. Gerade die Komplexität und Fülle der Daten, die in der digitalen Epidemiologie anfallen, erfordern es, neue Methoden zu entwickeln, um Strukturen in diesen Daten „sichtbar zu machen“. Wichtig ist hier auch, dass interaktive Visualisierungen es erlauben, verschiedene Perspektiven auf einen Datensatz bereitzustellen und schnell zwischen verschiedenen Perspektiven zu wechseln. Leider wird die Macht der interaktiven Visualisierung noch unterschätzt. Intuitive Visualisierungen werden oft zu Unrecht mit der attraktiven Aufbereitung von Daten verwechselt. Diese Fehleinschätzung folgt aus dem Missverständnis, dass es in der Wissenschaft in erster Linie darum geht, Hypothesen zu testen, als neue Hypothesen zu entwickeln, die Intuition für Systeme zu schärfen und Daten explorativ zu verstehen. Interaktive Visualisierungen sind ein Instrument, eine neue Technologie, die ähnlich wie Mikroskop oder Teleskop neue Einblicke in Daten vermitteln, die dann quantitative Analysen anregen, bei der Entwicklung neuer Hypothesen helfen und neue Studien motivieren können. Immer mehr Internetplattformen, Onlinedatenbanken und Softwarepakete werden entwickelt, bei denen interaktive Visualisierungen eine dominante Rolle spielen. Die Technologie D3js.org („data driven documents“; [34]) oder die Programmieroberfläche „Observable“ [35] sind sehr elegante und vielversprechende Methoden, um komplexe Daten im Browser effektiv darzustellen. 2017 wurde die Internetplattform „Complexity Explorables“ [36] ins Leben



**Abb. 6** ▲ Netzwerktheoretische Berechnung der möglichen Ausbreitungswege im Kontext der Ebolaepidemie in Westafrika im Jahr 2013. Ausgangspunkt ist hier der Flughafen in Freetown, Sierra Leone. Das Netzwerk repräsentiert die wahrscheinlichsten Ausbreitungswege über das weltweite Flugverkehrsnetz. Berechnet wurden diese Wege durch spezielle Algorithmen, die den gesamten weltweiten Flugverkehr berücksichtigen

gerufen, auf der auch Prinzipien und Zusammenhänge aus dem Bereich der Epidemiologie veranschaulicht werden.

### Maschinelles Lernen – künstliche Intelligenz

Die am meisten verwendeten Schlagworte im Kontext „digitale Epidemiologie“ sind „maschinelles Lernen“ und „künstliche Intelligenz“. Maschinelles Lernen umfasst eine Klasse von Computeralgorithmen, die besonders geeignet sind, in komplexen, hochdimensionalen, großen und/oder lückenhaften Daten entweder Strukturen zu erkennen oder aus Daten Vorhersagen zu generieren, die traditionellen, linearen statistischen Methoden verborgen bleiben [37]. Es geht also, grob gesprochen, um neue Regressions- oder Klassifizierungsmethoden. Die Algorithmen des maschinellen Lernens zeichnen aus, dass innere Parameter zwar ver-

änderlich sind und optimiert werden, aber typischerweise nicht interpretierbar sind. Vorhersagen oder Klassifizierungen werden mit Trainingsdaten bewertet und während des Lernprozesses die inneren Parameter des Algorithmus so adaptiert, dass immer bessere „Scores“ erzielt werden. Diese Scores werden auf den Trainingsdaten berechnet und sind je nach Anwendung ein Maß für den Erfolg der Klassifizierung bzw. der Vorhersagen. Besonders neuronale Netzwerke [38] zeigen große Erfolge bei der Klassifizierung komplexer Daten. Diesen Algorithmen dient die Struktur biologischer neuronaler Netze, z. B. des visuellen Kortex, als Architekturvorlage mit hintereinandergeschalteten Lagen einzelner in-silico-simulierter Nervenzellen, deren laterale und vertikale synaptische Kopplungsstärken modifiziert werden können. Sogenannte Deep-Learning-Algorithmen sind nur dadurch ausge-

zeichnet, dass sie vergleichsweise viele Lagen von Neuronen implementieren [39]. Neuronale Netzwerke dieser Art werden schon seit den 1990er-Jahren verwendet und studiert [40]. Aber erst jetzt stehen Prozessorarchitekturen zur Verfügung, die effiziente Anwendungen erlauben. Außerdem stehen erst jetzt Datensätze mit dem notwendigen Volumen zur Verfügung. Die Erfolge neuronaler Netze bei der Bild-, Text- oder Spracherkennung sind frappierend und nur in Ansätzen verstanden [41]. Sie finden auch in der medizinischen Forschung immer zahlreichere Anwendungen. So können mittlerweile maschinelle Lernverfahren Melanome deutlich besser als Spezialisten identifizieren [42].

Eine vielversprechende Anwendung im Bereich Public Health ist die Früherkennung von Signalen in Surveillance-Daten. Andere momentan entwickelte Anwendungen sind die Erkennung neuer



Antibiotikaresistenzkombinationen im Rahmen der am Robert Koch-Institut (RKI) gesammelten Daten der Antibiotika-Resistenz-Surveillance [43]. Auch auf dem Gebiet nichtübertragbarer Krankheiten spielt maschinelles Lernen eine immer wichtigere Rolle. So werden in verschiedenen Projekten am RKI die im Rahmen der Studie zur Gesundheit von Kindern und Jugendlichen in Deutschland (KiGGS-Studie; [44]) gewonnenen Daten mithilfe maschineller Lernverfahren untersucht, um ggf. versteckte Zusammenhänge zu identifizieren.

## Datenschutz

Ein ganz wichtiger Aspekt der digitalen Epidemiologie sind die Herausforderungen, die durch die hochaufgelöste Aggregation personalisierter Daten entstehen. Nahezu 100 % aller Menschen nutzen mittlerweile das Internet und mobile Endgeräte täglich. Damit werden Internetunternehmen wie Apple, Google, Facebook oder Amazon persönliche Daten mit höchster Auflösung preisgegeben. Es wachsen gesellschaftlich das Bewusstsein, dass die eigenen Daten einen Wert haben, und der Unmut darüber, dass Endnutzer diese Daten großen Unternehmen „schenken“. Die Aussage: „If it's free, you are the product!“, trifft bei den genannten Technologien und Dienstleistungen in besonderem Maße zu.

Erfreulicherweise wächst mit gleicher Geschwindigkeit die Erkenntnis, dass Datenschutz gewährleistet werden muss, um Persönlichkeitsrechte nicht zu verletzen. Allerdings wird auch hier der Begriff Datenschutz wieder recht vage verwendet. Es wird oftmals nicht erkannt oder verstanden, welche Daten geschützt werden sollten. Die digitale Epidemiologie ist als Wissenschaft auf diese neuen Daten angewiesen. Es überrascht daher nicht, dass mittlerweile verschiedene Methoden entwickelt werden, um wissenschaftliche Erkenntnisse aus den personenspezifischen Verhaltensdaten zu gewinnen, wobei gleichzeitig ein hoher Grad an Datensicherheit gewährleistet wird. Mittlerweile gibt es interessante gemeinnützige Projekte, die Infrastruktur zur Verfügung stellen, damit Menschen „ihre persönlichen Daten“ spenden kön-

nen. Ein prominentes Beispiel ist die Internetplattform [www.openhumans.org](http://www.openhumans.org) [45]. Hier kann jeder Nutzer bzw. Datenspende genau auswählen, welche wissenschaftlichen Projekte die eigenen Daten nutzen dürfen. Die Portale sorgen dafür, dass die Spender auch in regelmäßigen Abständen über die Forschungsergebnisse informiert werden und darüber, welchen Wert die Datenspende hatte. Die Grundidee ist hier, die eigenen Daten für das Gemeinwohl zu spenden. Diese Philosophie ist sicher eine der vielversprechendsten und wird, so ist zu hoffen, weiterentwickelt, um die Datengrundlage der digitalen Epidemiologie zu erweitern und letztendlich die Gesundheit der Menschen zu verbessern.

## Ausblick

Für das nächste Jahrzehnt ist zu erwarten, dass sich die noch junge Disziplin „digitale Epidemiologie“ als fundamentaler Zweig der Epidemiologie sowohl in der Grundlagenforschung als auch in der anwendungsorientierten Forschung etablieren wird. Bei der Nutzung und Zusammenführung der für die Epidemiologie oftmals neuartigen Daten müssen dringend fachübergreifende Maßnahmen zur Einschätzung der Qualität und Validität entwickelt werden und diese bei der Nutzung der Daten dann auch in die Bewertung der Ergebnisse einfließen. Die größte Herausforderung wird es sein, die Methoden zeitnah in die Ausbildung von Epidemiologen und Wissenschaftlern im Bereich Public Health zu integrieren. In verschiedenen anderen natur-, sozial- und lebenswissenschaftlichen Disziplinen werden schon seit einigen Jahren die neuen Methoden systematisch in die universitäre Ausbildung integriert, aber im internationalen Vergleich doch eher zaghafte. Für den Fortschritt in den Gesundheitswissenschaften ist es wünschenswert, dass auch die Lehrenden den Chancen der neuen Methoden offen begegnen und im Sinne der nächsten Generation von Wissenschaftlern darauf achten, dass die Ausbildung im Bereich Informatik, Datenanalyse und Computerprogrammierung schon sehr früh als

zentrales Element jedes Curriculums etabliert wird.

## Korrespondenzadresse

**Prof. Dr. Dirk Brockmann**  
Robert Koch-Institut  
Nordufer 20, 13353 Berlin, Deutschland  
[dirk.brockmann@hu-berlin.de](mailto:dirk.brockmann@hu-berlin.de)

**Funding.** Open Access funding provided by Projekt DEAL.

## Einhaltung ethischer Richtlinien

**Interessenkonflikt.** D. Brockmann gibt an, dass kein Interessenkonflikt besteht.

Für diesen Beitrag wurden von den Autoren keine Studien an Menschen oder Tieren durchgeführt. Für die aufgeführten Studien gelten die jeweils dort angegebenen ethischen Richtlinien.

**Open Access** Dieser Artikel wird unter der Creative Commons Namensnennung 4.0 International Lizenz veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden.

Die in diesem Artikel enthaltenen Bilder und sonstiges Drittmaterial unterliegen ebenfalls der genannten Creative Commons Lizenz, sofern sich aus der Abbildungslegende nichts anderes ergibt. Sofern das betreffende Material nicht unter der genannten Creative Commons Lizenz steht und die betreffende Handlung nicht nach gesetzlichen Vorschriften erlaubt ist, ist für die oben aufgeführten Weiterverwendungen des Materials die Einwilligung des jeweiligen Rechteinhabers einzuholen.

Weitere Details zur Lizenz entnehmen Sie bitte der Lizenzinformation auf <http://creativecommons.org/licenses/by/4.0/deed.de>.

## Literatur

1. C. elegans Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282:2012–2018
2. International Human Genome Sequencing Consortium (IHGSC) (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431:931–945
3. Heather JM, Chain B (2016) The sequence of sequencers: the history of sequencing DNA. *Genomics* 107:1–8
4. Schuster SC (2008) Next-generation sequencing transforms today's biology. *Nat Methods* 5:16–18
5. Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nat Biotechnol* 26:1135–1145
6. Neumann B, Bender JK, Maier BF et al (2019) Combining clinical epidemiology, NGS-based analysis and modelling approaches to reveal transmission dynamics of vancomycin-resistant enterococci in a high risk population within

- a tertiary care hospital. *PLoS Comput Biol* (under review)
7. Salathé M, Bengtsson L, Bodnar TJ et al (2012) Digital epidemiology. *PLoS Comput Biol* 8:e1002616
  8. Pastor-Satorras R, Castellano C, Van Mieghem P, Vespignani A (2015) Epidemic processes in complex networks. *Rev Mod Phys* 87:925
  9. Freifeld CC, Mandl KD, Reis BY, Brownstein JS et al (2008) Healthmap: global infectious disease monitoring through automated classification and visualization of internet media reports. *J Am Med Inform Assoc* 15:150–157
  10. ProMED International Society for Infectious Diseases (2019) Undiagnosed disease, bean - Bangladesh: (RS). <http://www.promedmail.org/>. Zugegriffen: 22. Okt. 2019
  11. International Society of Travel Medicine (2019) Geosentinel—the global surveillance network of the ISTM in partnership with the CDC. <https://www.istm.org/geosentinel>. Zugegriffen: 22. Okt. 2019
  12. OIE—World Organization for Animal Health (2019) Homepage. <https://www.oie.int/>. Zugegriffen: 22. Okt. 2019
  13. FAO—Food and Agriculture Organization of the United Nations (2019) Homepage. <http://www.fao.org/home/en/>. Zugegriffen: 22. Okt. 2019
  14. EuroSurveillance—Europe's journal on infectious disease surveillance, epidemiology, prevention and control (2019) Homepage. <https://www.eurosurveillance.org/>. Zugegriffen: 22. Okt. 2019
  15. Bhatt S, Gething PW, Brady OJ et al (2013) The global distribution and burden of dengue. *Nature* 496:7446
  16. Hadfield J, Megill C, Bell SM et al (2018) Next-strain: real-time tracking of pathogen evolution. *Bioinformatics* 1:4121–4323
  17. Yamayoshi S, Kawaoka Y (2019) Current and future influenza vaccines. *Nat Med* 25:212
  18. Paul MJ, Dredze M, Broniatowski D (2014) Twitter improves influenza forecasting. *PLoS Curr*. <https://doi.org/10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117>
  19. Salathé M, Khandelwal S (2011) Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. *PLoS Comput Biol* 7:e1002199
  20. Mønsted B, Lehmann S (2019) Algorithmic detection and analysis of vaccine-denialist. Sentiment clusters in social networks (arXiv:1905.12908)
  21. Colleoni E, Rozza A, Arvidsson A (2014) Echo chamber or the public sphere? predicting political orientation and measuring political homophily in twitter using big data. *J Commun* 64:317–332
  22. Barrat A, Cattuto C, Colizza V, Pinton J-F, Van den Broeck W, Vespignani A (2010) High resolution dynamical mapping of social interactions with active RFID. *PLoS One* 5:e11596
  23. Barrat A, Cattuto C, Szomszor M, Van den Broeck W, Alani H (2010) Social dynamics in conferences: analyses of data from the live social semantics application. In: Patel-Schneider PF et al (Hrsg) *The semantic web—ISWC 2010*. ISWC 2010. Lecture notes in computer science 6497. Springer, Berlin, Heidelberg.
  24. Szomszor M, Kostkova P, Cattuto C, Van den Broeck W, Barrat A, Alani H (2010) Providing enhanced social interaction services for industry exhibitors at large medical conferences
  25. Isella L, Stehlé J, Barrat A, Cattuto C, Pinton J-F, Van den Broeck W (2011) What's in a crowd? Analysis of face-to-face behavioral networks. *J Theor Biol* 271:166–180
  26. Barrat A, Cattuto C, Colizza V, Isella L, Rizzo C, Tozzi AE, Van den Broeck W (2010) Wearable sensor networks for measuring face-to-face contact patterns in healthcare settings
  27. Vanhems P, Barrat A, Cattuto C et al (2013) Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PLoS One* 8:e73970
  28. Stopczynski A, Sekara V, Sapiezynski P et al (2014) Measuring large-scale social networks with high resolution. *PLoS ONE* 9:e95978
  29. Holme P, Masuda N (2015) The basic reproduction number as a predictor for epidemic outbreaks in temporal networks. *PLoS ONE* 10:e120567
  30. Speidel L, Klemm Eguiluz VM, Masuda N (2016) Temporal interactions facilitate endemicity in the susceptible-infected-susceptible epidemic model. *New J Phys* 18:73013
  31. Valdano E, Valdano E, Ferreri L, Poletto C, Colizza V (2015) Analytical computation of the epidemic threshold on temporal networks. *Phys Rev X* 5:21005
  32. Soccaletti S, Latora V, Moreno Y, Chavez M, Hwang D-U (2006) Complex networks: structure and dynamics. *Phys Rep* 424:175–308
  33. Wu X, Liu Z (2008) How community structure influences epidemic spread in social networks. *Physica A* 387(2):623–630. <https://doi.org/10.1016/j.physa.2007.09.039>
  34. Data Driven Documents (2019) Homepage. <https://d3js.org/>. Zugegriffen: 19. Okt. 2019
  35. Observable (2019) Homepage. <https://observablehq.com/>. Zugegriffen: 19. Okt. 2019
  36. Complexity Explorables (2019) Homepage. <http://www.complexity-explorables.org/>. Zugegriffen: 19. Okt. 2019
  37. Kotsiantis SB, Zaharakis ID, Pintelas PE (2006) Machine learning: a review of classification and combining techniques. *Artif Intell Rev* 26:159–190
  38. Gurney K (1997) An introduction to neural networks
  39. Längkvist M, Karlsson L, Loutfi A (2014) A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognit Lett* 42:11–24
  40. Müller B, Reinhardt J, Strickland MT (1995) *Neural networks: an introduction*. Springer, Heidelberg
  41. Chakraborty S, Tomsett R, Raghavendra R et al (2017) Interpretability of deep learning models: a survey of results. 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation..
  42. Haenssle HA, Fink C, Schneiderbauer R (2018) Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol* 29(8):1836–1842. <https://doi.org/10.1093/annonc/mdy166>
  43. ARS – Antibiotika-Resistenz-Surveillance (2019) Webpräsenz. <https://ars.rki.de/>. Zugegriffen: 23. Okt. 2019
  44. KiGGS – Studie zur Gesundheit von Kindern und Jugendlichen in Deutschland (2019) Webpräsenz. <https://www.kiggs-studie.de>. Zugegriffen: 23. Okt. 2019
  45. Open Humans (2019) Homepage. <https://www.openhumans.org/>. Zugegriffen: 23. Okt. 2019