nature medicine

FOCUS | REVIEW ARTICLE

https://doi.org/10.1038/s41591-019-0727-5

Axes of a revolution: challenges and promises of big data in healthcare

Smadar Shilo[1,2,3,4], Hagai Rossman[1,2,4] and Eran Segal[1,2*]

Health data are increasingly being generated at a massive scale, at various levels of phenotyping and from different types of resources. Concurrent with recent technological advances in both data-generation infrastructure and data-analysis methodologies, there have been many claims that these events will revolutionize healthcare, but such claims are still a matter of debate. Addressing the potential and challenges of big data in healthcare requires an understanding of the characteristics of the data. Here we characterize various properties of medical data, which we refer to as 'axes' of data, describe the considerations and tradeoffs taken when such data are generated, and the types of analyses that may achieve the tasks at hand. We then broadly describe the potential and challenges of using big data in healthcare resources, aiming to contribute to the ongoing discussion of the potential of big data resources to advance the understanding of health and disease.

Health has been defined as "a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity"[1]. This definition may be expanded to view health not as a single state but rather as a dynamic process of different states in different points in time that together assemble a health trajectory[2]. The ability to understand the health trajectories of different people, how they would unfold along different pathways, how the past affects the present and future health, and the complex interactions between different determinants of health over time are among the most challenging and important goals in medicine.

Following technological, organizational and methodological advances in recent years, a new and promising direction has emerged toward achieving those goals: the analysis of large medical and biological datasets. With the rapid increase in the amount of medical information available, the term 'big data' has become increasingly popular in medicine. This increase is anticipated to continue as data from electronic health records (EHRs) and other emerging data sources such as wearable devices and multinational efforts for collection and storage of data and biospecimens in designated biobanks will expand.

Analyses of large-scale medical data have the potential to identify new and unknown associations, patterns and trends in the data that may pave the way to scientific discoveries in pathogenesis, classification, diagnosis, treatment and progression of disease. Such work includes using the data for constructing computational models to accurately predict clinical outcomes and disease progression, which have the potential to identify people at high risk and prioritize them for early intervention strategies[3], and to evaluate the influence of public health policies on 'real-world' data[4]. However, many challenges remain for the fulfillment of these ambitious goals.

In this Review, we first define big data in medicine and the various axes of medical data, and describe data-generation processes, more specifically considerations for constructing longitudinal cohorts for obtaining data. We then discuss data-analysis methods, the potential goals of these analyses and the challenges for achieving them.

Big data in medicine

The definition of 'big data' is diverse, in part because 'big' is a relative term. Although some definitions are quantitative, focusing on the volume of data needed for a dataset to be considered big[5], other definitions are qualitative, focusing on the size or complexity of data that are too large to be properly analyzed by traditional data-analysis methods[6]. In this Review, we refer to 'big data' as qualitatively defined.

Medical data have unique features compared with big data in other domains[7]. The data may include administrative health data, biomarker data, biometric data (for example, from wearable technologies) and imaging, and may originate from many different sources, including EHRs, clinical registries, biobanks, the internet and patient self-reports[8]. Medical data can also be characterized and vary by states such as (i) structured versus unstructured (for example, diagnosis codes versus free text in clinical notes); (ii) patient-care-oriented versus research-oriented (for example, hospital medical records versus biobanks); (iii) explicit versus implicit (for example, checkups versus social media), and (iv) raw versus ascertained (data without processing versus data after standardization and validation processes).

Defining axes of data

Health data are complex and have several different properties. As these properties are quantitative, we can view them as 'axes' of the data. Some properties may be easy to quantify, such as the number of participants, the duration of longitudinal follow up, and the depth, which may be calculated as the number of different types of data being measured. Other properties may be more challenging to quantify, such as heterogeneity, which may be computed using various diversity indices[9]. In this context, healthcare data may be viewed as having the axes described below (Figs. 1 and 2b).

Number of participants (axis N). Sample size in an important consideration in every medical data source. In longitudinal cohorts, planning the desired cohort size—calculated on the basis of an estimate of the number of predefined clinical endpoints expected to occur during the follow-up period—is critical to reaching sufficient statistical power[10]. As a result, a study of rare disease trajectory before symptom onset would require a very large number of subjects and is often impractical. Retention rate is also important in determining the cohort size[11]. The main limitations for increasing sample size are the recruitment rate, and financial and organizational constraints.

[1]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel. [2]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel. [3]Pediatric Diabetes Unit, Ruth Rappaport Children's Hospital, Rambam Healthcare Campus, Haifa, Israel. [4]These authors contributed equally: Smadar Shilo, Hagai Rossman. *e-mail: eran.segal@weizmann.ac.il

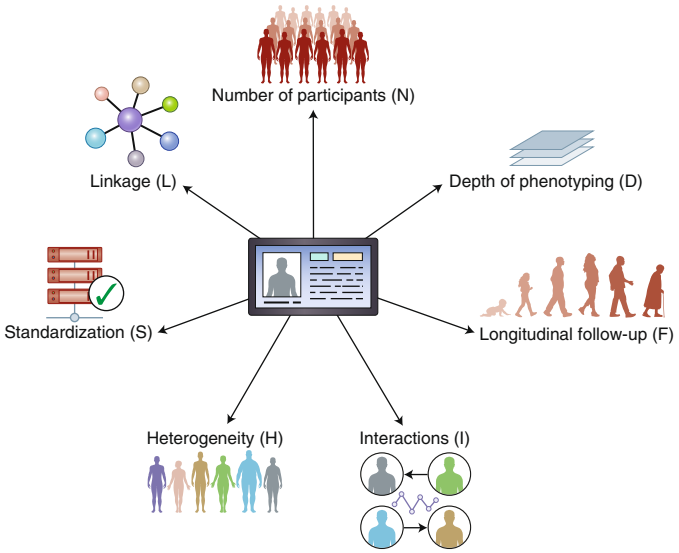


Fig. 1 | The different axes of health data. The complexity of large health datasets can be represented by distinct axes, each encompassing a quantifiable property of the data.

Depth of phenotyping (axis D). Medical data may range from the molecular level up to the level of social interactions among subjects. It may be focused on one specific organ or system in the body (such as the immune system) or may be more general and contain information about the entire body (as with total-body magnetic resonance imaging).

At the molecular level, data may be obtained by a variety of methods that analyze a diverse array of 'omics' data, which broadly represents the information contained within a person's genome and its biological derivatives. Omics data may include transcriptional, epigenetic, proteomic and metabolomic data[12]. Another rich source of omic-level information is the human microbiome, the collective genome of trillions of microbes that reside in the human body[13].

Additional phenotypes that may be obtained include demographics and socioeconomic factors (for example, ethnicity and material status), anthropometrics (for example, weight and height measurements), lifestyle habits (for example, smoking, exercise, nutrition), physiome or continuous physiological measurements (for example, blood pressure, heart rhythm and glucose measurements, which can be measured by wearable devices), clinical phenotyping (for example, diagnoses, medication use, medical imaging and procedure results), psychological phenotyping and environmental phenotyping (for example, air pollution and radiation level by environmental sensors that connect with smartphones). Diverse data types pose an analytical challenge, as their processing and integration requires in-depth technical knowledge about how these data were generated, the relevant statistical analyses, and the quantitative and qualitative relationship of different data types[14].

In the construction of a prospective cohort, the choice of the type and the depth of information to measure is challenging and depends on many considerations. Each test should be evaluated on the basis of its relevance, reliability and required resources. Relevance relies on other epidemiological studies that found significant associations with the studied health outcomes. Reliability includes selecting methods that pass quality testing, including calibration, maintenance, ease of use, training, monitoring and data transfer. Resources include both capital and recurrent costs.

Additional considerations include finding the right balance between exploiting known data types (such as genomic information) and exploring new types of data (such as new molecular assays) that have not been previously studied for the scientific question and are therefore more risky but may lead to new and exciting discoveries (hence exploration versus exploitation). It is also important to consider that the rapid acceleration of newer and cheaper technologies for data processing, storage and analysis will hopefully enable measurements of more data types and for larger cohorts as time progresses. One striking example is the cost of DNA sequencing, which decreased over one-million-fold in the past two decades[15]. Another consideration is the possibility that the mechanisms sought, and the answers to the scientific questions, depend on components that we cannot currently measure; therefore, considering which biospecimens to store for future research is also important.

Longitudinal follow-up (axis F). Longitudinal follow-up includes the total duration of follow-up, the time intervals between data points (or follow-up meetings, in the case of longitudinal cohorts), and the availability of different data types in each point. Long-term follow-up allows observation of the temporal sequence of events.

It has been hypothesized that the set point of several physiological and metabolic responses in adulthood is affected by stimulus or insults during the critical period of embryonic and fetal life development, a concept known as 'fetal programming'[16]. For example, associations between low birthweight and type 2 diabetes mellitus, coronary heart disease and elevated blood pressure have been demonstrated[17]. Therefore, for full exploration of disease mechanisms, the follow-up period should ideally be initiated as early as possible, with data collection starting from the preconception stage, followed by the pregnancy period, delivery, early and late childhood, and adulthood (hence the 'from pre-womb to tomb' approach)[18]. Although such widespread information is rarely available in most data sources, large longitudinal studies that recruit women at pregnancy are emerging, such as The Born in Guangzhou Cohort Study[19] and the Avon Longitudinal Study of Parents and Children[20].

Another important consideration in longitudinal cohorts is adherence of the participants to follow-ups. Selection bias owing to loss to follow-up may negatively affect the internal validity of the study[21]. For example, the UKBiobank was criticized as having selection bias because of the low response rate by participants (5.5%)[22]. Disadvantaged socioeconomic groups, including ethnic minorities, are more likely to drop out and thus possibly bias the results. It is therefore important to consider the effect of various retention strategies on different subpopulations in longitudinal studies, specifically for studies with a long follow-up period[11]. To increase adherence to follow-ups, incentives are sometimes used. For example, the Genes for Good study uses incentives such as interactive graphs and visualizations of survey responses, as well as personal estimates of genetic ancestry, for participant retention[23].

Interactions between subjects included in the data (axis I). The ability to connect each subject in the data to other people who are related to him or her is fundamental to the ability to explore mechanisms of disease onset and progression, and gene-environment interactions. Such relations may be genetic, which would allow calculation of the genetic distance between different people, or environmental, such as identifying people who share the same household, workplace, neighborhood or city. Intentional recruitment of subjects with genetic or environmental interactions increases the power to answer these scientific questions. One example is twin cohorts, such as the Finnish Twin Cohort[24] or recruitment of family triads of mothers, fathers and their offspring, such as The Norwegian Mother and Child Cohort Study[25]. Of note, recruitment of genetically related people or people from the same environment may result in decreased heterogeneity and diversity of the cohort.

Heterogeneity and diversity of the cohort population (axis H). Including factors such as age, sex, race, ethnicity, disability status,

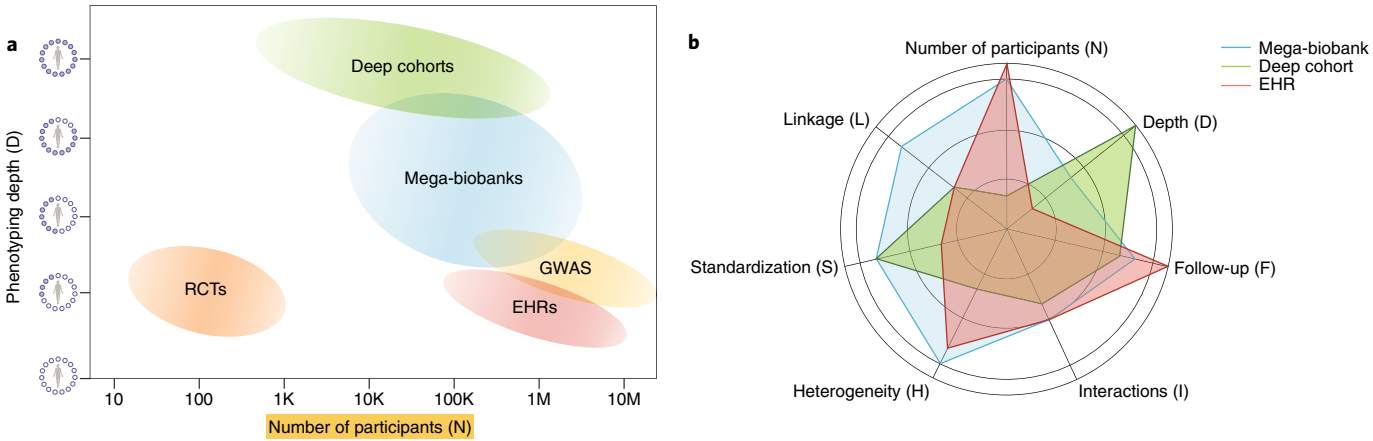


Fig. 2 | Tradeoffs between axes of data. a, Examples of various types of data-generating cohorts and their crude placement on the axes of phenotyping depth versus number of participants. GWAS, genome-wide association studies; RCTs, randomized controlled trials. **b**, Axes values for three types of cohorts (key): mega-biobanks, deep cohorts and EHRs.

socioeconomic status, educational level and geographic location is important. The process of selecting a cohort that will fully represent the real-world population is challenging. Challenges arise from a variety of historical, cultural, scientific and logistical factors, as the inclusion process involves several steps: selection of a subject for inclusion in the study, consent of the subject, and selection of the subject data to be analyzed by the study researchers. Sampling bias may arise at each of these steps, as different factors may affect them[26]. One example is volunteer bias, as it has been shown that people who are willing to participate in studies may be systematically different from the general population[27].

However, high heterogeneity in the study population and inclusion of disadvantaged socioeconomic groups are important for generalization of the results to the entire population. Medical research of under-represented minorities and people of non-European ancestry is often lacking in many fields[28]. One of the most prominent examples of this is in genetics, in which the vast majority of participants in genome-wide association studies are of European descent[29]. Many other fundamental studies in medicine have included only a relatively homogenous population. For example, the original Framingham Heart Study[30], which included residents of the city of Framingham, Massachusetts, and the Nurses' Health Study[31], which included registered American nurses, were relatively homogeneous in environmental exposures and education level, respectively. Thus, although many important studies were based on these cohorts, the question of whether their conclusions apply to the general population remains open[32]. Current studies such as the All of Us Research Program define heterogeneity as one of their explicit goals, with more than 80% of the participants recruited so far being from historically under-represented groups[33].

Nonetheless, increasing the heterogeneity of the study population (for example, by including participants of a young age) may increase the variability in the phenotype tested and decrease the anticipated rate of clinical endpoints expected to occur during the study period, and therefore will require a larger sample size to reach significant results.

Standardization and harmonization of data (S). Health data may come from many disparate data sources. Using these sources to answer desired clinical research questions requires comparing and analyzing these sources concurrently. Thus, harmonizing data and maintaining a common vocabulary are important. Data can be either collected in a standardized way (for example, ICD-9 diagnoses, structured and validated questionnaires) or can be categorized at a later stage by standard definitions.

Standardizing medical data into a universal format will enable collaborations across multiple countries and resources[34,35]. For example, the Observational Health Data Sciences and Informatics initiative is an international collaborative effort to create open-source unified common data models from a transformed large network of health databases[34]. This enables a significant increase in sample size and in heterogeneity of data, as shown in a recent study that examined the effectiveness of second-line treatment of type-2 diabetes, using data made available by the Observational Health Data Sciences and Informatics initiative from 246 million patients from multiple countries and cohorts[36].

Another interesting solution is to characterize and standardize descriptions of datasets in a short identification document that will accompany them, a concept described as 'datasheets for datasets'. Such a document will include the characteristics, motivations and potential biases of the dataset[37].

Linkage between data sources (L). The ability to link different data sources and thereby retrieve information on a specific person from several data sources is also of great value. For example, UKBiobank data are partially linked to existing health records, such as those from general practice, hospitals and central registries[38]. Linking EHRs with genetic data collected in large cohorts enables the correlation of genetic information with hundreds to thousands of phenotypes identified by the EHR[39].

For this linkage to be possible, each person should be issued a unique patient identifier that will apply across databases. However, mostly due to privacy and security concerns, unique patient identifiers are currently not available[40]. For tackling this, two main approaches have been suggested. The first is to create regulation and legislative standards to ensure the privacy of the participants. The second is to give patients full ownership of their own information and thereby allow them to choose whether they permit linkage to some or all of their medical information. For example, Estonia was the first country to give its citizens full access to their EHRs[41,42]. The topic of data ownership is debatable and has been discussed elsewhere[43,44].

Additional aspects of medical data have been previously described as part of the FAIR principles for data management: findable, accessible, interoperable and reusable. The data should be (i) findable, specifically registered or indexed in a searchable resource, because knowing which data exist is not always easy; (ii) accessible, as access to data by the broad scientific community is important for it to reaching its full scientific potential; (iii) interoperable, with a formal and accessible applicable language for knowledge

representation, which is also a part of the standardization axis described above; and (iv) reusable, which includes developing tools for scalable and replicable science, a task that requires attention and resources[45].

How is big data generated?

Longitudinal population studies and biobanks represent two sources of big data. Whereas much of the medical data available for analysis is passively generated in healthcare systems, new forms of biobanks, which actively generate data for research purposes, have been emerging in recent years. Biobanks were traditionally defined as collections of various types of biospecimens[46]. This definition has been expanded to "a collection of biological material and the associated data and information stored in an organized system, for a population or a large subset of a population"[47]. Biobanks have increased in variety and capacity, combining different types of phenotyping data; this has created rich data resources for research[48]. Unlike traditional, single-hypothesis-driven studies, these rich datasets try to address many different scientific questions. The prospective nature of these studies is especially important, because the effects of different factors on disease onset can be analyzed.

Although the concept of mega biobanks[49] is not well defined in the literature, it can be viewed qualitatively as biobanks that integrate many of the data axes mentioned above at a broad scale and includes data measured on large sample sizes (axis N) together with deep phenotyping of each subject (axis D) for a long follow-up period (axis F), collected and stored with standardization (axis S), and allowing interactions between participants (axis I) and with external sources (axis L) to be studied. Prominent examples of these include UKBiobank[50], All of Us Research[33], Kadoorie biobank[51], Million Veteran program[49] and Mexico City study[52], as well as others. A comprehensive survey of existing biobanks is presented in the review in ref. [26].

'Deep cohorts': a tradeoff between axes

In the construction of a biobank or a longitudinal cohort, each of the axes of data mentioned above has to be carefully assessed, as each has its costs and benefits. Limited research resources dictate an inherent tradeoff between different axes, and the ideal dataset that measures everything on everybody is unattainable. One necessary tradeoff is between the scale of the data gathered (axis N) and the depth of the data (axis D). For example, EHRs can contain medical information on millions of people but rarely include any molecular phenotypes or lifestyle assessments. Another example is '*N*-of-1 trials'. These could be used as a principled way to design trials for personalized medicine[53] or run a deep multidimensional profile of carefully selected subjects[54].

Medium-sized cohorts of hundreds or tens of thousands of people represent an interesting operating point, as they allow collection of full molecular and phenotypic data on a large enough population and thus enable the study of a wide variety of scientific questions. We can term such cohorts 'deep cohorts'.

Since delicate disease patterns may be detected only when the data include a deep enough phenotyping (axis D) of a sufficient sample size (*N*), deep cohorts that apply the most-advanced technologies to phenotype, collect and analyze data from medium-sized cohorts may have an immense scientific potential. For example, we previously collected data for a cohort of over 1,000 healthy people and deeply phenotyped it for genetics, oral and gut microbiome, immunological markers, serum metabolites, medical background, bodily physical measures, lifestyle, continuous glucose levels and dietary intake. This cohort allowed us to study many scientific questions, such as the inter-person variability in post-meal glucose responses[55], the ability to predict human traits from microbiome data, factors that shape the composition of the microbiome[56], and associations between microbial genomic structural variants and host disease risk factors[57]. We are following this cohort longitudinally and expanding its number of participants by tenfold, as well as adding new types of assays, with the goal of identifying molecular markers for disease with diagnostic, prognostic and therapeutic value. Other examples of medium-sized cohorts include the University College London-Edinburgh–Bristol Consortium, which performs large-scale, integrated genomics analyses and includes roughly 30,000 subjects[58], and the Lifelines cohort, which deeply phenotyped subset of ~1,000 of its ~167,000-subject cohort for microbiome, genetics and metabolomics[59].

The other axes of medical data mentioned above also require financial resources. Therefore, planning a prospective cohort warrants careful consideration of these tradeoffs and utilization of cost-effective strategies. For example, both the duration of longitudinal follow-up, and the number and types of tests that are performed during follow-up visits (axis F) have financial costs. Increasing the heterogeneity of the cohort (axis H) may also come at a cost: in the All of Us Research Program, US National Institutes of Health funding was provided to support recruitment of community organizations to increase the cohort's racial, ethnic and geographic diversity[33]. Additional tradeoffs are very likely to come up when collecting data, some of which we discussed above in the individual axes sections. The tradeoffs between different axes of medical data and specifically between scale (axis N) and depth (axis D) are presented in Fig. 2.

Numerous additional challenges exist in the construction of a large longitudinal cohort[26]. Many of the challenges that arise from the collection, storage, processing and analysis of any medical data (as discussed in the 'Potential and challenges' subsection below) are amplified as the scale and the complexity of the data increase. In most cases, specialized infrastructure and expertise are needed to overcome these challenges, as the generation of new cost-effective high-throughput data requires expertise in different fields. In addition, many research applications emanating from these sources of data are interdisciplinary in nature. This presents an organizational challenge in creating collaborations between clinicians and data scientists, and in educating physicians to understand and apply tools for large-scale data sources.

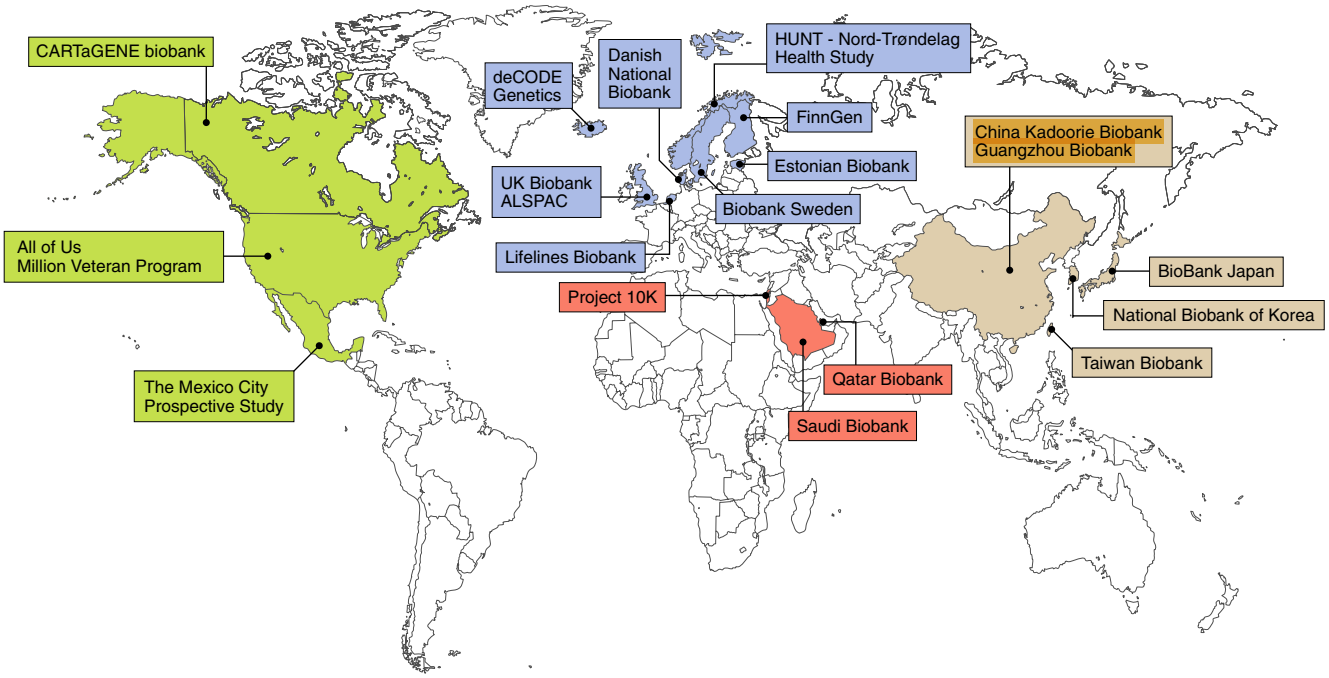
Ensuring participant compliance with the study protocol is also essential for ensuring scientific merit of the data. Several examples of this include fasting before blood tests and accurate logging of daily nutrition and activities in a designated application[55]. Compliance assessment by itself can also be challenging, as it often relies on self-reporting by participants. Finally, maintaining public trust and careful consideration of legal and ethical issues, especially those regarding privacy and de-identification of study participants, are crucial to the success of these studies[60–63].

Constructing a biobank requires considerable resources and, as a result, biobanks are much harder to establish in low- and middle-income countries. As a result, these populations remain under-represented and under-studied. The geographical distribution of the main biobanks worldwide is presented in Fig. 3.

How is big data analyzed?

How can utilization of these massive datasets achieve the potential of medical data analyses? How can we bridge the gap between the collected data, and our understanding and knowledge of human health? The answer to these questions can be broadly described by the common term 'data science'. Data science has been defined by as being segregated into three distinct forms of analysis tasks: description, prediction and counterfactual prediction[64]. This distinction holds true for medical data of any type and scale, and helps with the temptation to conflate different types of questions about analysis of the data[65]. These tasks can be defined and used as described below.

Descriptive analysis. Descriptive analysis can be broadly defined as "using data to provide a quantitative summary of certain features



Location	Biobank	N (goal)
Canada	CARTaGENE biobank[119]	43,000
USA	All of Us[33] Million Veteran Program[49]	1,000,000 > 600,000
Mexico	The Mexico City Prospective Study[52]	150,000
Iceland	deCODE Genetics	500,000
UK	UK Biobank[38] Avon Longitudinal Study of Parents and Children (ALSPAC)[20]	500,000 > 15,000
Netherlands	Lifelines Biobank[120]	> 167,000
Denmark	Danish National Biobank[121]	
Norway	HUNT - Nord-Trøndelag Health Study[122]	125,000
Sweden	Biobank Sweden	
Finland	FinnGen	500,000
Estonia	Estonian Biobank[123]	52,000
Israel	Project 10K	10,000
Saudi Arabia	Saudi Biobank	200,000
Qatar	Qatar Biobank[124]	60,000
China	China Kadoorie Biobank[51] Guangzhou Biobank[125]	> 500,000 30,000
Japan	BioBank Japan[126]	200,000
Korea	National Biobank of Korea[127]	500,000
Taiwan	Taiwan Biobank[128]	200,000

Fig. 3 | Global distribution of several biobanks and cohorts. Geographical distribution of the main biobanks and cohort studies that are currently collecting and analyzing health data. Websites: deCODE Genetics, https://www.decode.com/; Biobank Sweden, https://biobanksverige.se/english/research/; FinnGen, https://www.finngen.fi/en/finngenresearchprojectisanexpeditiontothefrontierofgenomicsandmedicine; Project 10K, http://www.weizmann.ac.il/sites/Project10K/; Saudi Biobank, https://kaimrc.med.sa/?page_id=1454.

of the world"[64]. A few examples include retrospective analyses of the dynamics of body mass index (BMI) in children over time in order to define the age at which development of sustained obesity occurs[66], and correlation of the differences in the normal body temperature of different people and mortality[67]. Descriptive analysis approaches are useful for unbiased exploratory study of the data and for finding interesting patterns in the data, which may lead to testable hypotheses.

Prediction analysis. Prediction analysis aims to learn a mapping from a set of inputs to some outcome of interest, such that the mapping can later be used to predict the outcome from the inputs

in a different unseen set. It is thus applied in settings in which there is a well-defined task. Prediction analysis holds the potential for improving disease diagnostic and prognostic (as discussed in the 'Potential and challenges' subsection below). Of note, the ability to construct accurate predictive models is heavily reliant on the availability of big data. Perhaps the most striking and famous examples are the recent advances in neural networks[68], which rely heavily on data at a large-enough scale and on advances in computing infrastructure; this enables the construction of prediction models.

Algorithmic advances in images, sequences and text processing have been phenomenal in recent years, riding on the wave of big data and deep learning methods. Taking the field of image recognition as an example, one of the most important factors for the phenomenal recent success was the creation and curation of a massive image dataset known as 'ImageNet'[69]. One hope is that the accumulation of similarly large, ascertained datasets in the medical domain can advance healthcare tasks at a magnitude similar to that of the change in image-recognition tasks. Prominent examples are Physionet[70] and the MIMIC dataset[71], which have been instrumental in advancing machine-learning efforts in health research[72]. These data have been used for competitions and as a benchmark for quite a few years, and are increasing in size and depth. Reviews on the potential of machine learning in health are provided in refs. [73–76].

One particularly promising direction of deep learning combined with massive datasets is that of 'representation learning'[77]; that is, finding the appropriate data representation, especially when the data are high-dimensional and complex. Healthcare data are usually unstructured and sparse, and can be represented by different techniques, based on domain knowledge to fully automated approaches. The representation of medical data with all of its derivatives (clinical narratives, examination reports, lab tests and others) should be in a form that will enable machine-learning algorithms to learn models with the best performance from it. In addition, the data representation may transform the raw data into a form that allows human interpretability with the appropriate model design[78].

Counterfactual prediction. One major limitation of any observational study is its inability to answer causal questions, as observational data may be heavily confounded and contain other limiting flaws[79]. These confounders may lead to high predictive power of a model that is driven by a variety of health processes rather than a true physiological signal[80]. Although proper study design and use of appropriate methods tailored to the use of observational data for causal analysis[81–84] may alleviate some of these issues, this remains an important open problem. One promising direction that uses some of the data collected at large scale to tackle causal questions is Mendelian randomization[85]. Studies involving large-scale genetic data and phenotypes combined with prior knowledge may have some ability to estimate causal effects[86]. Counterfactual prediction thus aims to construct models that address limiting flaws inherent to observational data for inferring causality.

Potential and challenges

The promise of medical big data depends on the ability to extract meaningful information from large-scale resources in order to improve the understanding of human health. We discussed some of the potentials and challenges of medical data analysis above. Additional broad categories that can be transformed by medical data include those discussed below.

Disease diagnosis, prevention and prognosis. The use of computational approaches to accurately predict future onset of clinical outcomes has the potential for early diagnoses, and either prevention or decrease in the occurrence of disease in both community and hospital settings. As some clinical outcomes have well-established modifiable risk factors, such as cardiovascular disease[87], prediction of these outcomes may enable early, cost-effective and focused preventive strategies for high-risk populations in the community setting. In the hospital setting, and specifically in intensive care units, early recognition of life-threatening conditions enables an earlier response from the medical team, which may lead to better clinical outcomes. Numerous prediction models have been developed in recent years. One recent example is the prediction of inpatient episodes of acute kidney injury[88]. Another example is the prediction of sepsis, as the early administration of antibiotics and intravenous fluids is considered crucial for the management of sepsis[89]. Several machine-learning-based sepsis-prediction algorithms have been published[90], and a randomized clinical trial demonstrated the beneficial real-life potential of this approach, decreasing patient length of stay in the hospital and in-hospital mortality[91].

Similarly, the same approach can be used to predict the prognosis of a patient with a given clinical diagnosis. Identifying subgroups of patients who are most likely to deteriorate or develop a certain complications of the disease can enable targeting of these patients and the use of strategies such as more frequent follow-up schedule, changes in medication regime or a shift from traditional care to palliative care[92].

Devising a clinically useful prediction model is challenging for several reasons. The predictive model should be continuously updated, accurate, well-calibrated and delivered at the individual level with adequate time for early and effective intervention by clinicians. It should help identify the population in which an early diagnostic or prognostic will benefit a patient. Therefore, prediction of unpreventable or incurable disease is of less immediate use, although such models may be clinically relevant in the future, as new therapeutics and prevention strategies emerge. Another important consideration is model interpretability, which includes understanding of the mechanism by which the model works; that is, model transparency or post hoc explanations of the model. Defining the very notion of interpretability is not so straightforward, and it may mean different things[93]. Finally, a predictive model should strive to be cost-effective and broadly applicable. A model based on existing information in EHR data is much more economical than a model based on costly molecular measurement.

The real-life success of a predictive model depends both on its performance and on the efficacy of prevention strategies that physicians can apply when they receive the information output by the model. One of the concerns about the real-life implementation of prediction models is that it will eventually result in over-diagnosis. Through the use of highly sensitive technologies, it is possible to detect abnormalities that would either disappear spontaneously or have a very slow and clinically unimportant progression. As a result, it is possible that more people will be unnecessarily labeled as being at high risk[94]. Another concern is algorithmic bias, which may be introduced in many ways. For example, it has been shown that an algorithm that is widely used by health systems exhibits racial bias[95]. Thus far, very few predictive models have been assessed in a real-life setting, and more studies are needed to validate the clinical utility of these tools per each specific clinical endeavor.

Modeling disease progression. Chronic diseases often progress slowly over a long period of time. Whereas some medical diagnoses are currently based on predefined thresholds, such as a hemoglobin A1C percentage of 6.5% or above for the diagnosis of diabetes mellitus[96], or a BMI of 30 kg/m^2 or above for the diagnosis of obesity (https://www.who.int/topics/obesity/en/), these diseases may be viewed as a continuum, rather than as a dichotomic state. Modeling the continuous nature of chronic diseases and progression over time is often challenging due to many reasons, such as incompleteness and irregularity of data, and heterogeneity of patient comorbidities and medication usage. Large-scale deep phenotyping of subjects can help overcome these challenges and allow a better understanding of

disease progression[97]. Notably, this view of disease as a continuum may allow the study of early stages of disease in healthy cohorts, without confounders such as medications and treatments, provided that the disease markers are well defined, measured and span enough variation in the studied population. Diabetes (diagnosed via hemoglobin A1C percentage), obesity (diagnosed via BMI) and hyperlipidemia (diagnosed via cholesterol values) are good examples in which this can be done, and may lead to the definition of disease risk scores for various diseases.

Genetic and environmental influence on phenotypes. The information on genetic and environmental exposures collected in biobanks combined with data on health outcomes can also lead to many discoveries on the effects of genetic and environmental determinants on disease onset and progression[98]—that is, nature versus nurture—and quantification of the magnitude of each of these determinants[99]. Despite many advances in genetic research over the past decades, major challenges such as small sample sizes and low population heterogeneity still remain[29]. This has led to the emergence of a new approach that uses EHR-driven genomic research, which combines data available in the EHR and phenotypic characterizations, and enables calculation of the effect size of a genetic variant not for one disease or trait but for all diseases simultaneously, also called a 'phenome-wide association study'[100,101]. However, the use of large-scale data sources also raises challenges in standards for defining disease and in efforts to extract characteristics of patients from EHRs, which is not always a straightforward task. To do so, one needs to incorporate medical knowledge on the data-generation process and validate the algorithms of extraction from raw data (https://www.phekb.org/).

Target identification. The development of new drugs is a very complex process, with over 90% of the chemical entities tested not making it to the market[102]. This process starts with identification of disease-relevant phenotypes and includes basic research, target identification and validation, lead generation and optimization, preclinical testing, phased clinical trials in humans, and regulatory approval (Fig. 4). Target identification, defined as 'identifying drug targets for a disease', and target validation, defined as 'demonstrating an effect of perturbation of the target on disease outcomes and related biomarkers', are essential parts in drug discovery and development.

The traditional pharmaceutical industry's screening process for the identification of new drug targets is costly and long, and includes activity assays, in which the compounds are tested through the use of high-throughput methods, based on interaction with the relevant target proteins or selected cell lines, and low-throughput methods, run on tissues, organs or animal models. This traditional screening method is characterized by a high dropout rate, with thousands of failures per one successful drug candidate[102]. Animal models are often used for these tasks, but they have a substantial disadvantage in the development of new drugs because their limited congruence with many human diseases severely affects their translational reliability[103].

There is thus a great need for new approaches to drug development. Human multi-omics data and physiological measurements at scale from deeply phenotyped cohorts is one such direction and is considered one of the most promising potentials of analyzing big data in medicine, as humans themselves will serve as the future model organisms[104,105]. First, analysis of large-scale health data may identify new, unknown associations[106] and therefore may allow the discovery of new biomarkers and novel drug targets, such as by mapping existing genetic-association findings to drug targets and compounds[107]. Second, analysis of biological and medical data may be used to evaluate the chances of success of drugs discovered and tested on animal models before the costly and time-consuming

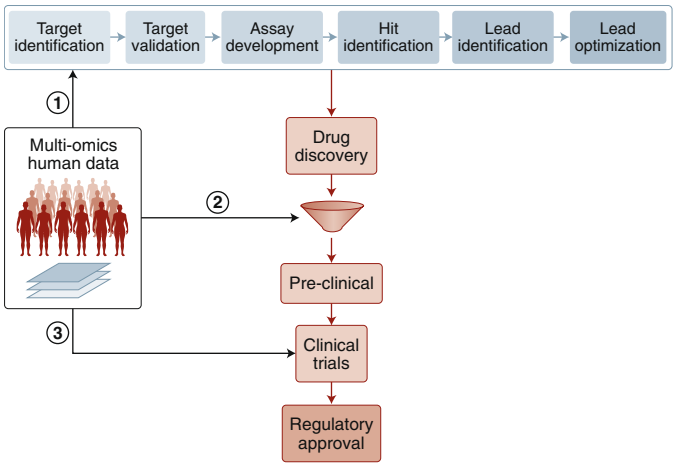


Fig. 4 | Using human-based omics data in drug development. Utilization of large-scale human multi-omics data in the process of drug development may aid in the following: (1) identification of new drug targets; (2) evaluation of drug candidates that were identified by animal models using human data before preclinical and clinical trials and; (3) identification of therapeutic targets with a well-established safety profile, which may be considered for a direct evaluation in clinical trials in humans.

stages of preclinical and clinical trials. Third, potential therapeutic interventions discovered via human data analysis with an established safety profile, such as nutritional modification or supplements and drugs with existing approval by the US Food and Drug Administration, may be considered for direct evaluation in human clinical trials (Fig. 4). Finally, human data can be used to investigate differences in drug response and potential side effects[104]. Since some drugs affect only a subset of the treated target patient population, using human data to distinguish responders from non-responders, and to prioritize responders for clinical trials, can have great utility. Analysis of large-scale human omics data therefore has the potential to accelerate drug development and reduce its cost. Indeed, it has been estimated that selecting targets with evidence from human genetics data may double the success rate of the clinical development of drugs[108].

Systematic analysis of large-scale data by various computational approaches can also be used to obtain meaningful interpretations for the repurposing of existing drugs[109]. For example, clinical information from over 13 years of EHRs that originated from a tertiary hospital has led to the identification of over 17,000 known drug–disease associations and to the identification of terbutaline sulfate, an anti-asthmatic drug, as a candidate drug for the treatment of amyotrophic lateral sclerosis[110]. Another example is the use of publicly available molecular data for the discovery of new candidate therapies for inflammatory bowel disease[111].

Improvement of health processes. Big-data analysis can allow the investigation of health-policy changes and optimization of health processes[4]. It has the potential to reduce diagnostic and treatment errors, eliminate redundant tests[112] and provide guidance for better distribution of health resources[113]. Realizing the potential of this direction requires close interaction with medical organizations in order to map the existing processes, understand the clinical implications, and decide on the desired operating points, tradeoffs and costs of mis- and over-diagnoses.

Disease phenotyping. Phenotyping of disease and health and the study of variation between people represent another potential of studying rich and novel types of data. For example, we previously characterized the variation between healthy people in response to

food, based on deep phenotyping of a 1,000-person cohort that included, to our knowledge, the first large-scale continuous glucose monitoring and gut microbiota profiling of healthy people[55].

Another potential is to refine current phenotyping of disease. For example, there have been attempts to refine the classification of type 2 diabetes and find subgroups from available data[97,114]. Another example is Parkinson's disease, for which recent advances in genetics, imaging and pathologic findings coupled with observed clinical variability, have profoundly changed the understanding of the disease. Parkinson's disease is now considered to be a syndrome rather than a single entity, and the International Parkinson and Movement Disorders Society have commissioned a task force for the redefinition of this disease[115–117].

Precision medicine. Analysis of big data in health that takes into account individual variability in omics data, environment and lifestyle factors may facilitate the development of precision medicine and novel prevention and treatment strategies[118]. However, caution should be taken, with careful assessments of how much of the change observed in the phenotype tested is due to variability within people[53]. It is not obvious that many of the medical questions of interest will be answered through big datasets. Historically, small and well-designed experiments were the primary drivers of medical knowledge, and the burden of showing a change in this paradigm is now put on new methodologies.

Conclusion

Big data in medicine may provide the opportunity to view human health holistically, through a variety of lenses, each presenting an opportunity to study different scientific questions. Here we characterized health data by several axes that represent different properties of the data. The potential scientific value of collecting large amounts of health data on human cohorts has recently been recognized, with a rapid rise in the creation of large-scale cohorts aiming to maximize these axes. However, since maximizing each axis requires both resources and effort, it is inevitable that some axes come at the expense of others. Analysis of big data in health has many challenges and is in some sense a double-edged sword. On one hand, it provides a much wider perspective on states of health and disease, but on the other hand, it provides the temptation to delve into the details of molecular descriptions that may miss the big picture (as in the 'seeing the whole elephant' analogy). In addition, real-world evidence that it will translate into improved quality of care is currently lacking. However, the potential to improve healthcare is still immense, especially as patients' conditions and medical technologies become more and more complex over time. With the collection of more deeply phenotyped large-scale data, many scientific questions about disease pathogenesis, classification, diagnosis, prevention, treatment and prognosis can be studied and can potentially lead to new discoveries that may eventually revolutionize medical practice.

Received: 18 October 2019; Accepted: 3 December 2019;
Published online: 13 January 2020

References

1. Grad, F. P. The Preamble of the Constitution of the World Health Organization. *Bull. World Health Organ.* **80**, 981 (2002).
2. Burton-Jeangros, C., Cullati, S., Sacker, A. & Blane, D. *A Life Course Perspective on Health Trajectories and Transitions* Vol. 4 pp. 1–18 (Springer, 2015); https://link.springer.com/chapter/10.1007/978-3-319-20484-0_1
3. Obermeyer, Z. & Emanuel, E. J. Predicting the future—big data, machine learning, and clinical medicine. *N. Engl. J. Med.* **375**, 1216–1219 (2016).
4. Benke, K. & Benke, G. Artificial intelligence and big data in public health. *Int. J. Environ. Res. Public Health* **15**, E2796 (2018).
5. Baro, E., Degoul, S., Beuscart, R. & Chazard, E. Toward a literature-driven definition of big data in healthcare. *BioMed. Res. Int.* **2015**, 639021 (2015).
6. Gligorijević, V., Malod-Dognin, N. & Pržulj, N. Integrative methods for analyzing big data in precision medicine. *Proteomics* **16**, 741–758 (2016).
7. Cios, K. J. & Moore, G. W. Uniqueness of medical data mining. *Artif. Intell. Med.* **26**, 1–24 (2002).
8. Rumsfeld, J. S., Joynt, K. E. & Maddox, T. M. Big data analytics to improve cardiovascular care: promise and challenges. *Nat. Rev. Cardiol.* **13**, 350–359 (2016).
9. Koopmans, R. & Schaeffer, M. Relational diversity and neighbourhood cohesion. Unpacking variety, balance and in-group size. *Soc. Sci. Res.* **53**, 162–176 (2015).
10. Gould, A. L. Planning and revising the sample size for a trial. *Stat. Med.* **14**, 1039–1051 (1995).
11. Booker, C. L., Harding, S. & Benzeval, M. A systematic review of the effect of retention methods in population-based cohort studies. *BMC Public Health* **11**, 249 (2011).
12. Mason, C. E., Porter, S. G. & Smith, T. M. Characterizing multi-omic data in systems biology. *Adv. Exp. Med. Biol.* **799**, 15–38 (2014).
13. Cho, I. & Blaser, M. J. The human microbiome: at the interface of health and disease. *Nat. Rev. Genet.* **13**, 260–270 (2012).
14. Palsson, B. & Zengler, K. The challenges of integrating multi-omic data sets. *Nat. Chem. Biol.* **6**, 787–789 (2010).
15. Check Hayden, E. Is the $1,000 genome for real? *Nature* https://www.nature.com/news/is-the-1-000-genome-for-real-1.14530 (2014).
16. Kwon, E. J. & Kim, Y. J. What is fetal programming?: a lifetime health is under the control of in utero health. *Obstet. Gynecol. Sci.* **60**, 506–519 (2017).
17. Barker, D. J. In utero programming of chronic disease. *Clin. Sci.* **95**, 115–128 (1998).
18. Topol, E. J. Individualized medicine from prewomb to tomb. *Cell* **157**, 241–253 (2014).
19. Qiu, X. et al. The born in guangzhou cohort study (BIGCS). *Eur. J. Epidemiol.* **32**, 337–346 (2017).
20. Golding, J., Pembrey M., Jones, R. & ALSPAC Study Team. ALSPAC—The Avon Longitudinal Study of Parents and Children. *Paediatr. Perinat. Epidemiol.* **15**, 74–87 (2001).
21. Howe, C. J., Cole, S. R., Lau, B., Napravnik, S. & Eron, J. J. Jr. Selection bias due to loss to follow up in cohort studies. *Epidemiology* **27**, 91–97 (2016).
22. Swanson, J. M. The UK Biobank and selection bias. *Lancet* **380**, 110 (2012).
23. Brieger, K. et al. Genes for Good: engaging the public in genetics research via social media. *Am. J. Hum. Genet.* **105**, 65–77 (2019).
24. Kaprio, J. The Finnish Twin Cohort Study: an update. *Twin Res. Hum. Genet.* **16**, 157–162 (2013).
25. Magnus, P. et al. Cohort profile update: the Norwegian mother and child cohort study (MoBa). *Int. J. Epidemiol.* **45**, 382–388 (2016).
26. Beesley, L. J. et al. The emerging landscape of health research based on biobanks linked to electronic health records: existing resources, statistical challenges, and potential opportunities. *Stat. Med.* https://doi.org/10.1002/sim.8445 (2019).
27. Lau, B., Gange, S. J. & Moore, R. D. Interval and clinical cohort studies: epidemiological issues. *AIDS Res. Hum. Retroviruses* **23**, 769–776 (2007).
28. Chen, M. S. Jr., Lara, P. N., Dang, J. H. T., Paterniti, D. A. & Kelly, K. Twenty years post-NIH Revitalization Act: enhancing minority participation in clinical trials (EMPaCT): laying the groundwork for improving minority clinical trial accrual: renewing the case for enhancing minority participation in cancer clinical trials. *Cancer* **120**, 1091–1096 (2014).
29. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).
30. Mahmood, S. S., Levy, D., Vasan, R. S. & Wang, T. J. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *Lancet* **383**, 999–1008 (2014).
31. Colditz, G. A., Manson, J. E. & Hankinson, S. E. The Nurses' Health Study: 20-year contribution to the understanding of health among women. *J. Women's Health* **6**, 49–62 (1997).
32. Liao, Y., McGee, D. L., Cooper, R. S. & Sutkowski, M. B. How generalizable are coronary risk prediction models? Comparison of Framingham and two national cohorts. *Am. Heart J.* **137**, 837–845 (1999).
33. Denny, J. C. et al. The "All of Us" Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).
34. Hripcsak, G. et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud. Health Technol. Inform.* **216**, 574–578 (2015).
35. Rajkomar, A. et al. Scalable and accurate deep learning with electronic health records. *npj Digit. Med.* **1**, 18 (2018).
36. Vashisht, R. et al. Association of hemoglobin a1c levels with use of sulfonylureas, dipeptidyl peptidase 4 inhibitors, and thiazolidinediones in patients with type 2 diabetes treated with metformin: analysis from the observational health data sciences and informatics initiative. *JAMA Netw. Open* **1**, e181755 (2018).
37. Gebru, T. et al. Datasheets for datasets. *arXiv* 1803.09010 (2018).

38. Sudlow, C. et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
39. Wolford, B. N., Willer, C. J. & Surakka, I. Electronic health records: the next wave of complex disease genetics. *Hum. Mol. Genet.* **27**, R14–R21 (2018).
40. Weber, G. M., Mandl, K. D. & Kohane, I. S. Finding the missing link for big biomedical data. *J. Am. Med. Assoc.* **311**, 2479–2480 (2014).
41. Evans, R. S. Electronic health records: then, now, and in the future. *Yearb. Med. Inform.* **1**, S48–S61 (2016).
42. Tiik, M. & Ross, P. Patient opportunities in the Estonian electronic health record system. *Stud. Health Technol. Inform.* **156**, 171–177 (2010).
43. Montgomery, J. Data sharing and the idea of ownership. *New Bioeth.* **23**, 81–86 (2017).
44. Rodwin, M. A. The case for public ownership of patient data. *J. Am. Med. Assoc.* **302**, 86–88 (2009).
45. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
46. Hewitt, R. & Watson, P. Defining biobank. *Biopreserv. Biobank.* **11**, 309–315 (2013).
47. Organization for Economic Cooperation and Development. Glossary of Statistical Terms: Biobank. in *Creation and Governance of Human Genetic Research Databases (OECD)*. https://stats.oecd.org/glossary/detail.asp?ID=7220 (2006).
48. Kinkorová, J. Biobanks in the era of personalized medicine: objectives, challenges, and innovation: Overview. *EPMA J.* **7**, 4 (2016).
49. Gaziano, J. M. et al. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).
50. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
51. Chen, Z. et al. China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int. J. Epidemiol.* **40**, 1652–1666 (2011).
52. Tapia-Conyer, R. et al. Cohort profile: the Mexico City Prospective Study. *Int. J. Epidemiol.* **35**, 243–249 (2006).
53. Senn, S. Statistical pitfalls of personalized medicine. *Nature* **563**, 619–621 (2018).
54. Garrett-Bakelman, F. E. et al. The NASA Twins Study: A multidimensional analysis of a year-long human spaceflight. *Science* **364**, eaau8650 (2019).
55. Zeevi, D. et al. Personalized nutrition by prediction of glycemic responses. *Cell* **163**, 1079–1094 (2015).
56. Rothschild, D. et al. Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210–215 (2018).
57. Zeevi, D. et al. Structural variation in the gut microbiome associates with host health. *Nature* **568**, 43–48 (2019).
58. Shah, T. et al. Population genomics of cardiometabolic traits: design of the University College London-London School of Hygiene and Tropical Medicine-Edinburgh-Bristol (UCLEB) Consortium. *PLoS One* **8**, e71345 (2013).
59. Tigchelaar, E. F. et al. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open* **5**, e006772 (2015).
60. Cohen, I.G. & Mello, M.M. Big data, big tech, and protecting patient privacy. *J. Am. Med. Assoc.* **322**, 1141–1142 (2019).
61. Price, W. N. II & Cohen, I. G. Privacy in the age of medical big data. *Nat. Med.* **25**, 37–43 (2019).
62. Tutton, R., Kaye, J. & Hoeyer, K. Governing UK Biobank: the importance of ensuring public trust. *Trends Biotechnol.* **22**, 284–285 (2004).
63. Kaufman, D. J., Murphy-Bollinger, J., Scott, J. & Hudson, K. L. Public opinion about the importance of privacy in biobank research. *Am. J. Hum. Genet.* **85**, 643–654 (2009).
64. Hernán, M. A., Hsu, J. & Healy, B. A second chance to get causal inference right: a classification of data science tasks. *Chance* **32**, 42–49 (2019).
65. Shmueli, G. To Explain or to Predict? *Stat. Sci.* **25**, 289–310 (2010).
66. Geserick, M. et al. Acceleration of BMI in early childhood and risk of sustained obesity. *N. Engl. J. Med.* **379**, 1303–1312 (2018).
67. Obermeyer, Z., Samra, J. K. & Mullainathan, S. Individual differences in normal body temperature: longitudinal big data analysis of patient records. *Br. Med. J.* **359**, j5468 (2017).
68. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
69. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).
70. Goldberger, A. L. et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* **101**, E215–E220 (2000).
71. Johnson, A. E. W. et al. MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 160035 (2016).
72. Wang, S. et al. MIMIC-Extract: a data extraction, preprocessing, and representation pipeline for MIMIC-III. *arXiv* 1907.08322 (2019).
73. Ching, T. et al. Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **15**, 20170387 (2018).
74. Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L. & Ranganath, R. Opportunities in machine learning for healthcare. *arXiv* 1806.00388 (2018).
75. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
76. Rajkomar, A., Dean, J. & Kohane, I. Machine learning in medicine. *N. Engl. J. Med.* **380**, 1347–1358 (2019).
77. Bengio, Y., Courville, A. & Vincent, P. Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1798–1828 (2013).
78. Weng, W.H. & Szolovits, P. Representation learning for electronic health records. *arXiv* 1909.09248 (2019).
79. Dickerman, B. A., García-Albéniz, X., Logan, R. W., Denaxas, S. & Hernán, M. A. Avoidable flaws in observational analyses: an application to statins and cancer. *Nat. Med.* **25**, 1601–1606 (2019).
80. Agniel, D., Kohane, I. S. & Weber, G. M. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *Br. Med. J.* **361**, k1479 (2018).
81. Hernán, M. A. & Robins, J. M. Using big data to emulate a target trial when a randomized trial is not available. *Am. J. Epidemiol.* **183**, 758–764 (2016).
82. Pearl, J. *Causality* (Cambridge University Press, 2009).
83. Johansson, F., Shalit, U. & Sontag, D. Learning representations for counterfactual inference. *arXiv* 1605.03661 (2016).
84. Dickerman, B. A., García-Albéniz, X., Logan, R. W., Denaxas, S. & Hernán, M. A. Avoidable flaws in observational analyses: an application to statins and cancer. *Nat. Med.* **25**, 1601–1606 (2019).
85. Smith, G. D. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32**, 1–22 (2003).
86. Hu, P., Jiao, R., Jin, L. & Xiong, M. Application of causal inference to genomic analysis: advances in methodology. *Front. Genet.* **9**, 238 (2018).
87. Yusuf, S. et al. Modifiable risk factors, cardiovascular disease, and mortality in 155 722 individuals from 21 high-income, middle-income, and low-income countries (PURE): a prospective cohort study. *Lancet* https://doi.org/10.1016/S0140-6736(19)32008-2 (2019).
88. Tomašev, N. et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* **572**, 116–119 (2019).
89. Rivers, E. et al. Early goal-directed therapy in the treatment of severe sepsis and septic shock. *N. Engl. J. Med.* **345**, 1368–1377 (2001).
90. Calvert, J. S. et al. A computational approach to early sepsis detection. *Comput. Biol. Med.* **74**, 69–73 (2016).
91. Shimabukuro, D. W., Barton, C. W., Feldman, M. D., Mataraso, S. J. & Das, R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. *BMJ Open Respir. Res.* **4**, e000234 (2017).
92. Avati, A. et al. Improving palliative care with deep learning. *BMC Med. Inform. Decis. Mak.* **18**, 122 (2018).
93. Lipton, Z. C. The mythos of model interpretability. *Commun. ACM* **61**, 36–43 (2018).
94. Vogt, H., Green, S., Ekstrøm, C. T. & Brodersen, J. How precision medicine and screening with big data could increase overdiagnosis. *Br. Med. J.* **366**, l5270 (2019).
95. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
96. American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diabetes Care* **36**, S67–S74 (2013)..
97. Udler, M. S. et al. Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. *PLoS Med.* **15**, e1002654 (2018).
98. Young, A. I., Benonisdottir, S., Przeworski, M. & Kong, A. Deconstructing the sources of genotype-phenotype associations in humans. *Science* **365**, 1396–1400 (2019).
99. Lakhani, C. M. et al. Repurposing large health insurance claims data to estimate genetic and environmental contributions in 560 phenotypes. *Nat. Genet.* **51**, 327–334 (2019).
100. Kohane, I. S. Using electronic health records to drive discovery in disease genomics. *Nat. Rev. Genet.* **12**, 417–428 (2011).
101. Phelan, M., Bhavsar, N. & Goldstein, B. A. Illustrating informed presence bias in electronic health records data: how patient interactions with a healthsystem can impact inference. *eGEMs* **5**, 22 (2017).
102. Brodniewicz, T. & Grynkiewicz, G. Preclinical drug development. *Acta Pol. Pharm.* **67**, 578–585 (2010).
103. Breyer, M. D. Improving productivity of modern-day drug discovery. *Expert Opin. Drug Discov.* **9**, 115–118 (2014).
104. FitzGerald, G. et al. The future of humans as model organisms. *Science* **361**, 552–553 (2018).

105. Matthews, H., Hanison, J. & Nirmalan, N. "Omics"-informed drug and biomarker discovery: opportunities, challenges and future perspectives. *Proteomes* **4**, 28 (2016).
106. Reshef, D. N. et al. Detecting novel associations in large data sets. *Science* **334**, 1518–1524 (2011).
107. Finan, C. et al. The druggable genome and support for target identification and validation in drug development. *Sci. Transl. Med.* **9**, eaag1166 (2017).
108. Nelson, M. R. et al. The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856–860 (2015).
109. Pushpakom, S. et al. Drug repurposing: progress, challenges and recommendations. *Nat. Rev. Drug Discov.* **18**, 41–58 (2019).
110. Paik, H. et al. Repurpose terbutaline sulfate for amyotrophic lateral sclerosis using electronic medical records. *Sci. Rep.* **5**, 8580 (2015).
111. Dudley, J. T. et al. Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Sci. Transl. Med.* **3**, 96ra76 (2011).
112. Xu, S. et al. Prevalence and predictability of low-yield inpatient laboratory diagnostic tests. *JAMA Netw. Open* **2**, e1910967 (2019).
113. Einav, L., Finkelstein, A., Mullainathan, S. & Obermeyer, Z. Predictive modeling of U.S. health care spending in late life. *Science* **360**, 1462–1465 (2018).
114. Ahlqvist, E. et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *Lancet Diabetes Endocrinol.* **6**, 361–369 (2018).
115. Thenganatt, M. A. & Jankovic, J. Parkinson disease subtypes. *JAMA Neurol.* **71**, 499–504 (2014).
116. Lawton, M. et al. Developing and validating Parkinson's disease subtypes and their motor and cognitive progression. *J. Neurol. Neurosurg. Psychiatry* **89**, 1279–1287 (2018).
117. Berg, D. et al. Time to redefine PD? Introductory statement of the MDS Task Force on the definition of Parkinson's disease. *Mov. Disord.* **29**, 454–462 (2014).
118. Collins, F. S. & Varmus, H. A new initiative on precision medicine. *N. Engl. J. Med.* **372**, 793–795 (2015).
119. Awadalla, P. et al. Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics. *Int. J. Epidemiol.* **42**, 1285–1299 (2013).
120. Scholtens, S. et al. Cohort Profile: LifeLines, a three-generation cohort study and biobank. *Int. J. Epidemiol.* **44**, 1172–1180 (2015).
121. Christensen, H., Nielsen, J. S., Sørensen, K. M., Melbye, M. & Brandslund, I. New national Biobank of The Danish Center for Strategic Research on Type 2 Diabetes (DD2). *Clin. Epidemiol.* **4**, 37–42 (2012).
122. Krokstad, S. et al. Cohort profile: the HUNT study, Norway. *Int. J. Epidemiol.* **42**, 968–977 (2013).
123. Leitsalu, L. et al. Cohort profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int. J. Epidemiol.* **44**, 1137–1147 (2015).
124. Al Kuwari, H. et al. The Qatar Biobank: background and methods. *BMC Public Health* **15**, 1208 (2015).
125. Jiang, C. Q. et al. An overview of the Guangzhou biobank cohort study-cardiovascular disease subcohort (GBCS-CVD): a platform for multidisciplinary collaboration. *J. Hum. Hypertens.* **24**, 139–150 (2010).
126. Nagai, A. et al. Overview of the BioBank Japan Project: study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).
127. Lee, J.-E. et al. National Biobank of Korea: quality control programs of collected-human biospecimens. *Osong Public Health Res. Perspect.* **3**, 185–189 (2012).
128. Lin, J.-C., Fan, C.-T., Liao, C.-C. & Chen, Y.-S. Taiwan Biobank: making cross-database convergence possible in the Big Data era. *Gigascience* **7**, 1–4 (2018).

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to E.S.

Peer review information Joao Monteiro was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.