מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

Axes of a revolution

Document Version:
Accepted author manuscript (peer-reviewed)

Citation for published version:
Shilo, S, Rossman, H & Segal, E 2020, 'Axes of a revolution: challenges and promises of big data in healthcare', *Nature Medicine*, vol. 26, no. 1, pp. 29-38. https://doi.org/10.1038/s41591-019-0727-5

Total number of authors:
3

Digital Object Identifier (DOI):
10.1038/s41591-019-0727-5

Published In:
Nature Medicine

License:
Other

General rights
@ 2020 This manuscript version is made available under the above license via The Weizmann Institute of Science Open Access Collection is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognize and abide by the legal requirements associated with these rights.

How does open access to this work benefit you?
Let us know @ library@weizmann.ac.il

Take down policy
The Weizmann Institute of Science has made every reasonable effort to ensure that Weizmann Institute of Science content complies with copyright restrictions. If you believe that the public display of this file breaches copyright please contact library@weizmann.ac.il providing details, and we will remove access to the work immediately and investigate your claim.

(article begins on next page)

Download date: 26. Mar. 2025

Axes of a revolution: How big data will transform healthcare

S. Shilo, H. Rossman, E. Segal

Abstract

Health data is increasingly being generated at massive scales, 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 these claims are still a matter of much debate. Addressing the potentials and challenges of big data in healthcare requires an understanding of the characteristics of the data. Here, we characterize different axes of medical data, describe the considerations and tradeoffs taken when generating such data and the types of analysis that may achieve the tasks at hand. Our review aims to dissect and discuss these topics and to contribute to the ongoing discussion of the shift to big data resources and its potential in advancing our understanding of health and disease.

Introduction

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 individuals, 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 towards achieving these goals: the analysis of big data in healthcare. 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 (EHR) and other emerging data sources such as wearable devices and multinational efforts for collection and storage of data and bio-specimens in designated Biobanks will expand.

Analyses of large-scale medical data has the potential to unravel new and unknown associations, patterns and trends in the data that may pave the way to many scientific discoveries in the pathogenesis, classification, diagnosis, treatment and progression of diseases. These include constructing computational models utilizing the data in order to accurately predict clinical outcomes and disease progression which have the potential to identify individuals of high risk and

prioritize them for early intervention strategies [3] and evaluation of the influence of public health policies on real life data [4]. However, many challenges still remain for the fulfillment of these ambitious goals.

In this review, we will first define big data in medicine and different axes of medical data, followed by data generation processes, and more specifically considerations for constructing longitudinal cohorts for obtaining data. We will 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, partly due to the fact that "Big" is a relative term. While some definitions are quantitative, defining the volume of data needed for a dataset to be considered as "Big" [5], other definitions are qualitative, such as datasets in which the size or complexity of the data are too large to be properly analysed using traditional data analysis methods [6].

Medical data has unique features [7]. It may include administrative health data, biomarkers, biometric data (e.g., from wearable technologies) and imaging and may originate from many different sources, including EHRs, clinical registries, biobanks, the Internet and patient's self reported data [8]. Medical data can also be characterized and vary by states such as (1) structured vs. unstructured (e.g., diagnosis codes vs. free text in clinical notes); (2) Patient-care vs. research oriented (e.g., hospital medical records vs. biobanks); (3) Explicit vs. implicit (e.g., checkups vs. social media) and (4) Raw vs. ascertained (e.g., data without processing vs. data after standardization and validation processes).

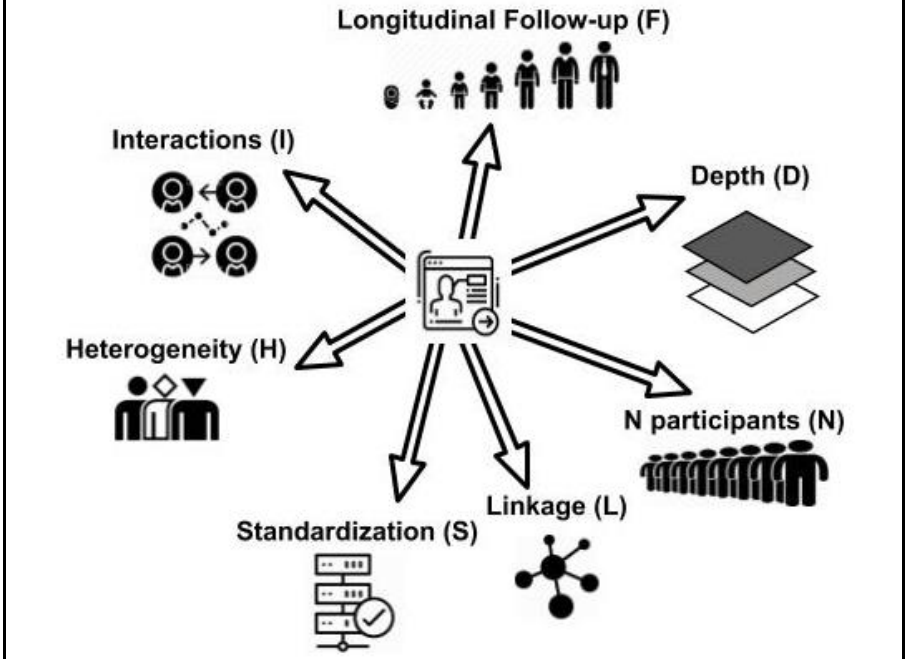

Figure 1: Axes of health data

Axes of data

Health data is complex, and can be viewed as having the following axes (Figure 1):

1. (N) **Number of participants**: Sample size in an important consideration in every medical data source. In longitudinal cohorts, planning of the desired cohort size, calculated based on the estimation of the number of predefined clinical endpoints expected to occur during the follow-up period, is critical to reach sufficient statistical power [9]. As a result, the study of rare disease trajectories even before their onset requires a very large number of individuals and is often impractical. Retention rate from the study is also an important factor in determining the cohort size [10]. The main limitations for increasing sample size are recruitment rate and other financial and organisational constraints.

2. (D) **Depth of phenotyping**: Medical data may range from the molecular level and scale up even to the level of social interactions among subjects. It may be focused on one specific organ or system in the body (e.g., the immune system) or be more general and contain information regarding the entire body (e.g., total body MRI imaging).

   On the molecular level, data may be obtained from a variety of experimental methods that analyze a diverse array of "omics" data, which broadly represents the information

contained within an individual's genome and its biological derivatives. Omic data may include transcriptional, epigenetic, proteomic and metabolomic data [11]. Another rich source of omic-level information is the human microbiome, the collective genome of trillions of microbes residing in our body [12].

Additional phenotypes which may be obtained include demographics and socioeconomic factors (e.g., ethnicity, material status), anthropometrics (e.g., weight and height measurements), lifestyle habits (e.g., smoking, exercise, nutrition), physiome or continuous physiological measurements (e.g., blood pressure, heart rhythm and glucose measurements, which can be measured by wearable devices), clinical phenotyping (e.g., diagnoses, medication use, medical imaging and procedures results), psychological phenotyping and environmental phenotyping (e.g., air pollution and radiation level by environmental sensors that connect with smartphones). Diverse data types poses 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 [13].

When constructing a prospective cohort, choosing the type and depth of information to measure is challenging and depends on many considerations. Each test should be evaluated based on 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 v.s. 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 1-million fold in the last two decades [14]. Another consideration is the possibility that the mechanisms that we are searching for, and the answers to our scientific questions, depend on components that we cannot currently measure, and therefore considering which biospecimies to store for future research is also of highly importance.

3. (F) **Longitudinal follow-up**: Long term follow-up allows observation of temporal sequence of events. We should consider the time intervals between different data points or

follow-up meetings in the case of longitudinal cohorts, the availability of different phenotypic data in each point, and the total duration of follow-up.

It was previously hypothesized that the set point of several physiological and metabolic responses in adulthood are affected by stimulus or insults during the critical period of embryonic and fetal life development, a concept known as "fetal programming" [15]. For example, associations between low birthweight and type 2 diabetes mellitus, coronary heart disease, and elevated blood pressure have been previously demonstrated [16]. Therefore, to be able to fully explore 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 "from pre-womb to tomb" approach [17]. Although such widespread information is rarely available in most data sources, large longitudinal cohorts that recruit women already at pregnancy are emerging, such as "The Born in Guangzhou Cohort Study" (BIGCS) [18] and the "Avon Longitudinal Study of Parents and Children" (ALSPAC) cohorts [19] .

Another important consideration in longitudinal cohorts is the adherence of the participants to follow-ups. Selection bias due to loss to follow-up may negatively affect the internal validity of the study [20]. For example, the UKBiobank was criticized as having a selection bias due to low response rate by participants (5.5%) [20]. Disadvantaged socio-economic 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 different retention strategies on different subpopulations in longitudinal studies, specifically in studies in which the follow-up period is long [10]. To increase adherence to follow-ups, incentives are sometimes utilized. For example, Genes for Good (GFG) uses incentives such as return of survey response summaries and genetic data including ancestry analysis to participants for continued participation and contribution of a saliva sample [21].

4. (I) **Interactions between individuals included in the data**: The ability to connect each individual in the data to other individuals who are related to him or her is fundamental for the ability to explore mechanisms of disease onset and progression and gene-environment interactions. Such relations may be genetic, allowing the calculation of the genetic distance between different individuals, or environmental, e.g., identifying individuals that share the same household, neighborhood or city. Intentional recruitment of individuals with genetic or environmental interactions increases the power to answer these scientific questions. One example is twin cohorts, such as "the Finnish Twin Cohort" [22] or recruitment of family triads of mothers, fathers and their offspring, such as "The Norwegian Mother and Child Cohort Study" (MoBa) [23]. Of note, recruitment of genetically related individuals or individuals from the same environmental may result in decreased heterogeneity and diversity of the cohort (see below).

5. (H) **Heterogeneity and diversity of the cohort population:** When analysing the heterogenicity of the data, factors such as age, sex, race, ethnicity, disability status, socioeconomic status, educational level, and geographic location should be considered. In longitudinal cohorts, the inclusion process involves several steps: a selection of a subject for inclusion in the study, the consent of the patient, and the selection of the subject data to be analysed by the study researchers. Sampling bias may arise from either one of these steps, as different factors may impact them [24]. For example, volunteer biases, as it was shown that individuals who are willing to participate in studies may be systematically different from the general population [25].

   High heterogeneity in the study population and inclusion of disadvantaged socio-economic groups are important for generalizing the results to the entire population and since medical research of these populations is often lacking [26]. For example, the Framingham Heart Study [27], which included residents of the city of Framingham, Massachusetts, and the Nurses' Health Study [28], which included registered American nurses, were relatively homogeneous in regard to environmental exposures and educational 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 [29]. Current studies such as the All-of-Us research program defined heterogeneity as one of their explicit goals, with more than 80% of the participants recruited to date originate from historically underrepresented groups [30].

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

6. (S) **Standardization and harmonization of data**: Health data may come from many disparate data sources. Using these sources to answer desired clinical research questions requires comparing and analysing these sources concurrently. As such, harmonizing data and maintaining a common vocabulary is important. Data can be either collected in a standardized way (e.g., ICD-9 diagnoses, structured and validated questionnaires) or be categorized in a later stage by standard definitions.

   Standardizing medical data into a universal format will enable collaborations across multiple countries and resources [31,32]. For example, the Observational Health Data Sciences and Informatics (OHDSI) initiative is an international collaborative effort to create open-source unified common data models from transformed large network of health databases [31]. This enables a significant increase in sample size and heterogeneity of data as shown in

a recent study that examined the effectiveness of second-line treatment of type 2 diabetes, and utilized data made available by this initiative from 246 million patients from multiple countries and cohorts [33].

7. (L) **Linkage between data sources:** The ability to link different data sources together, and thereby retrieve information on a specific individual from several data sources is also of great value. For example, The UKBiobank data is partially linked to existing health records, such as those from general practice, hospitals and central registries [34]. Linking EHR with genetic data collected in large cohorts enables the correlation of genetic information with hundreds to thousands phenotypes identified by the EHR [35].

   For this linkage to be possible, each person should be issued a unique patient identifier (UPI) that will apply to all databases. However, mostly due to privacy and security concerns, UPI's are currently not available [36]. To tackle this, two main approaches were suggested. The first is to create regulation and legislative standards to ensure the privacy of the participants. The second is to give the patients full ownership of their own information, thereby allowing them to choose whether they allow 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 [37,38]. The topic of data ownership is debatable and broadly discussed elsewhere [39,40].

   Additional aspects of medical data were previously described as part of the *FAIR* principles for data management. The data should be: (1) **Findable**, data registered or indexed in a searchable resource since knowing which data exists by itself is not always easy; (2) **Accessible**, as access to the data by the broad scientific community is important for reaching its full scientific potential; (3) **Interoperable,** a formal and accessible applicable language for knowledge representation, which is also a part of the standardization (S) axis described above, and; (4) **Reusable**, including developing tools for scalable and replicable science, a task that requires attention and resources [41].

Data generation

Longitudinal population studies and Biobanks

While much of the medical data available for analysis is passively generated by the health care systems, new forms of *biobanks*, which actively generate data for research purposes, are emerging in recent years. *Biobanks* were traditionally defined as collections of various types of biospecimens [42]. This definition was recently 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" [43]. *Biobanks* have increased in variety and capacity, combining different types of

phenotyping data, thus creating rich data resources for research [44]. Unlike traditional, single-hypothesis driven studies, these rich data sets try to address many different scientific questions. The prospective nature of these studies is especially important, because the effects of different risk factors can be analysed prior to disease onset.

While the concept of *"Mega Biobanks"* [45] is not well defined in the literature, we believe they can be viewed qualitatively as Biobanks which integrate many of the different data axes mentioned above at a broad scale and include data measured on large sample sizes (N) together with deep phenotyping of each individual (D) for a long follow-up period (F), collected and stored with standardization (S), and allowing interactions within participants (I) and with external sources (L) to be studied. Prominent examples include UKBiobank [46], All-of-us Research [30], Kadoorie biobank [47], Million Veteran program [45] and Mexico City study [48] as well as others. For a comprehensive survey of existing biobanks, see [24].

'Deep cohorts': A tradeoff between axes

When constructing 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 individuals, but rarely have any molecular phenotypes or lifestyle assessments. Medium size cohorts of thousands or tens of thousands of individuals represent an interesting operating point as they can collect full molecular and phenotypic data on a large enough population and thus study a wide variety of scientific questions. We can term such cohorts *'deep cohorts'*.

We believe that there is immense scientific potential for deep cohorts that apply the most advanced technologies to phenotype, collect, and analyze data from medium sized cohorts. For example, we previously collected 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 interpersonal variability in post-meal glucose responses [49–51], the ability to predict human traits from microbiome data, factors shaping the composition of the microbiome [50], and associations between microbial genomic structural variants and host disease risk factors [51]. We are following this cohort longitudinally and expanding its number of participants by 10-fold 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 size cohorts include the University College London-Edinburgh-Bristol (UCLEB) Consortium, which performs large-scale, integrated genomics analyses and includes roughly 30,000 individuals [52] and the Lifelines cohort which deeply phenotyped a ~1,000 subset of its ~167,000 person cohort for microbiome, genetics, and metabolomics [53] .

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, National Institutes of Health (NIH) funding was given to support recruitment of community organizations in order to increase the cohort racial, ethnic and geographic diversity [30]. Additional tradeoffs are very likely to come up when collecting data, some were 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) is presented in Figure 2.

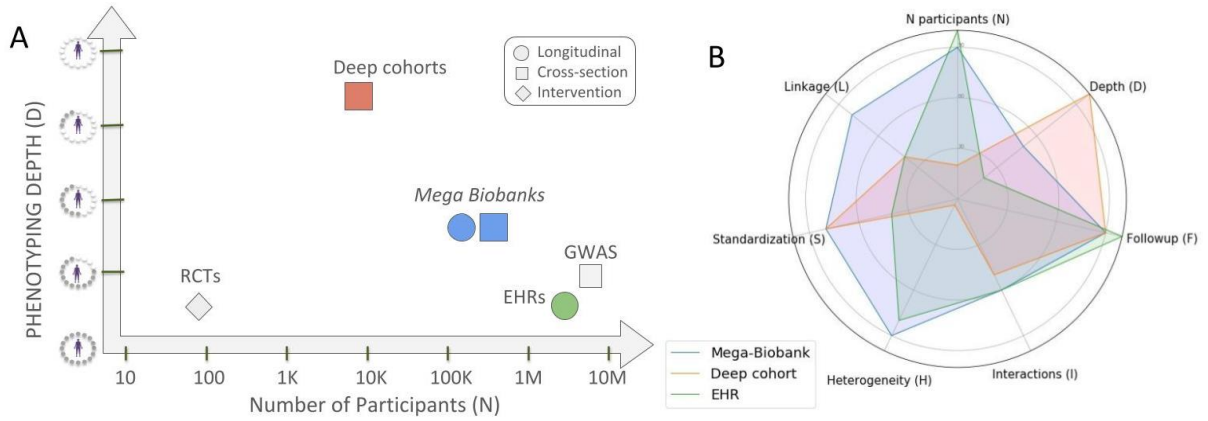

Figure 2: Tradeoffs between axes of data. A: Examples of types of data generating cohorts and their placement on the axes of phenotyping depth v.s. Number of participants. **B:** Axes values for three types of cohorts: Blue - Mega-Biobanks, Red: Deep cohorts, Green - EHRs.
Abbreviations : EHR- Electronic health record, GWAS - Genome wide association study, RCT- Randomized controlled trial.

Numerous additional challenges exist in the construction of a large longitudinal cohort [24]. Many of the challenges that arise in the collection, storage, processing and analysis of any medical data (see 'Potentials and Challenges' section below) are amplified as the scale and the complexity of the data increases. 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 their nature. This presents an organizational challenge in creating collaborations between clinicians and data scientists and educating physicians to understand and apply tools for large scale data sources.

Ensuring participants compliance to the study protocol is also essential for ensuring the scientific merit of the data. Several examples include fasting prior to blood tests [54] and accurate logging of daily nutrition and activities in a designated application [49]. Compliance assessment by itself can also be challenging, as it often relies on participants self reporting. 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 [55–58]; [55,56]

Constructing a Biobank requires many resources, and as a result, is much harder to establish in underdeveloped countries. As a result, these populations remain underrepresented and studied. The geographical distribution of the major biobanks that are currently being collected worldwide is presented in Figure 3.

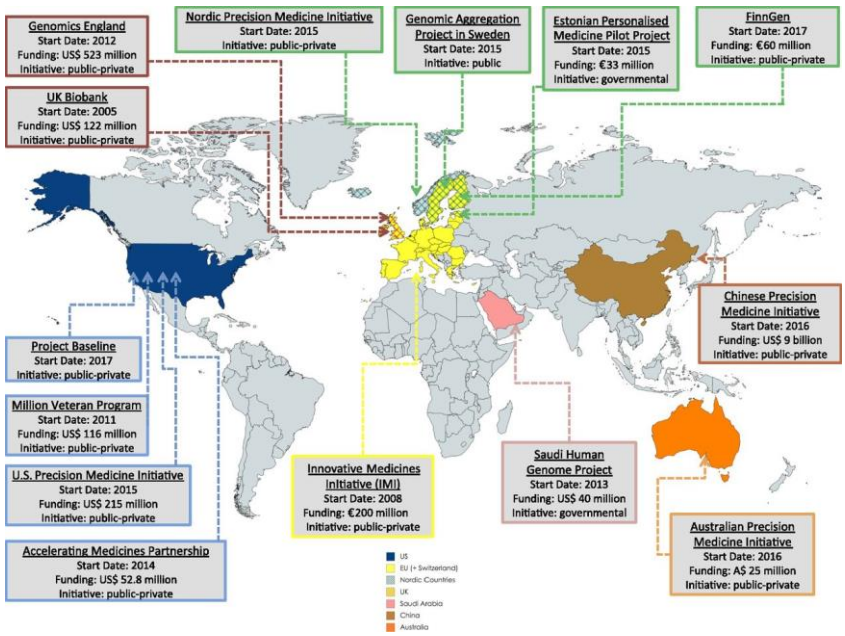

Figure 3: Global map of Mega-Biobanks

Data analysis

How do we utilize massive datasets to achieve the numerous potentials 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 was previously defined by Hern[59]n [59] as segregated into 3 distinct forms of analysis tasks: *Description, prediction,* and *counterfactual prediction*. This distinction holds true for medical data of any type and scale and helps us with the temptation to conflate different types of questions regarding analysis of the data [60].

These tasks can be defined and utilized as following:

1. *Descriptive* analysis can be broadly defined as "using data to provide a quantitative summary of certain features of the world" [59]. A few examples include retrospective analyses of the dynamics of Body Mass Index (BMI) in children over time in order to define the age in which development of sustained obesity occurs [61] and the correlation of the differences in normal body temperature within different individuals and mortality [62]. 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.

2. *Prediction* analysis aims to learn a mapping from a set of inputs to some outcome of interest where the mapping can later be used to predict the outcome from the inputs in a different set. It is thus applied in settings in which we have a well-defined task. Prediction analysis holds the potential for improving disease diagnostic and prognostic (see 'Potential and Challenges' section below). Of note, the availability of big data by itself may also be related to the predictive model success. Perhaps the most striking and famous examples are the recent advances in Neural Networks [63], which rely heavily on data at a large enough scale and on advances in computing infrastructure thus enabling 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 [64]. One hope is that accumulation of similar large, ascertained datasets in the medical domain can advance healthcare tasks in a similar magnitude to the change in image recognition tasks. Prominent examples are Physionet [65] and MIMIC dataset [66] which have been instrumental in advancing machine learning efforts in health research [67]. This data has been used for competitions and as a benchmark for quite a few years and is increasing in size and depth. See [68], [69], [70] and [71] for comprehensive reviews on the potentials of Machine Learning (ML) in health.

   One particularly promising direction of deep learning combined with massive datasets is that of *Representation Learning* [72], i.e. finding the appropriate data representation, especially when the data is 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, etc.) 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 transforms the raw data into a form which allows human interpretability with the appropriate model design [72,73].

3. *Counterfactual prediction:* One major limitation of any observational study is its inability to answer causal questions, as observational data may be heavily confounded [74]. These confounders may lead to high predictive power of a model being driven by health processes rather than a true physiological signal. While proper study design and use of appropriate methods tailored to the use of observational data for causal analysis [75–77] 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 (MR) [78]. Studies involving large scale genetic data and phenotypes combined with prior knowledge may have some ability to estimate causal effects [79]. Counterfactual prediction thus aims to construct causal models and address well-defined questions about causality.

Potential and Challenges

The promise of big medical data depends on our ability to extract meaningful information from large scale health data in order to improve our 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 includes the following:

1. *Disease diagnosis, prevention and prognosis*

The utilization of computational approaches to accurately predict future onset of clinical outcomes has the potential to early diagnose, and either prevent or decrease the occurrence of disease in both community and hospital settings. As some clinical outcomes have well established modifiable risk factors, such as cardiovascular disease [80], 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 the 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 [81]. Another example is sepsis prediction, as the early administration of antibiotics and intravenous fluids is considered crucial for sepsis management [82]. Several machine learning-based sepsis prediction algorithms have been published [83] and Randomised clinical trial (RCT) demonstrated the beneficial real life potential of this approach, decreasing patient's length of stay in the hospital and in-hospital mortality [84].

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 complication of the disease can enable targeting these patients and employ strategies such as more

frequent follow-ups schedule, changes in medication regime or shifting from traditional care to palliative care [85].

Devising a clinically useful prediction model is challenging due to 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 the clinicians. It should help identify population in which an early diagnostic or prognostic will benefit the patient. Therefore, prediction of unpreventable or incurable disease are of less immediate use, although such models may be clinically relevant in the future, as new therapeutics and prevention strategies will emerge. Another important consideration is model interpretability which includes the understanding of the mechanism by which the model works, i.e. model transparency or post-hoc explanations of the model. Surprisingly, defining the very notion of interpretability is not so straightforward and may mean different things [86]. Finally, a predictive model should strive to be cost effective and applicable broadly. A model which is based on existing information in the EHR data is much more economic than a model based on costly molecular measurement.

The real-life success of a predictive model depends both on its performance and the efficacy of prevention strategies that physicians can apply when they receive the information outputted by the model. One of the concerns regarding the real life implementation of prediction models is that it will eventually result in overdiagnosis. By using 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 labelled as being at high risk [87]. To date, very few predictive models were assessed in real life setting and more studies are needed to validate the clinical utility of these tools per each specific clinical endeavor.

2. *Modeling Disease progression*

Chronic diseases often progress slowly over a long period of time. While some medical diagnoses are currently based on predefined thresholds, such as Hemoglobin A1C% (HbA1C%) 6.5% or above for the diagnosis of diabetes mellitus [88] or a BMI of 30 Kg/m^2 or above for the diagnosis of obesity [89], these diseases may be viewed as a continuum, rather than a dichotomic state. Modeling the continuous nature of chronic diseases, and its progression over time is often challenging due to many reasons, such as incompleteness and irregularity of the data and heterogeneity of the patient comorbidities and medication usage. Large scale deep phenotyping of individuals can help overcome these challenges and allow a better understanding of disease progression [90]. Notably, this view of disease as a continuum may allow study of early stages of diseases 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 (via HbA1C%), obesity (via BMI), and cardiovascular disease (via cholesterol and other established risk factors) are good examples where this can be done, and may lead to the definition of disease risk scores (DRS) for various diseases.

3. *Genetic and environmental influences 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 for disease onset and progression [91], i.e. “nature v.. nurture”, and to quantify the magnitude of each of these determinants [92]. While many advances occurred in genetic research in the past decades, major challenges such as small sample sizes and low of population heterogeneity still remain [93]. This has led to an emergence of a new approach using EHR-driven genomic research (EDGR) which combines data available in the EHR and the phenotypic characterizations and enable calculating the effect size of a genetic variant not for one disease or trait but for all diseases simultaneously, also called phenome-wide association study (PheWAS) [94],[95]. However, the use of large scale data sources also raises challenges in standards for defining disease and 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 generating process and validate algorithms of extraction from the raw data [96].

4. *Target identification*

Development of new drugs is a very complex process, with over 90% of the chemical entities tested not making it to the market [97]. This process starts with identification of disease-relevant phenotypes and includes basic research, target identification and validation, lead generation and optimisation, pre-clinical testing, phased-clinical trials in humans, and regulatory approval (Figure 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 screening process for identification of new drug targets is costly and long, and includes activity assays, in which the compounds are tested using 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 drop-out rate, with thousands of failures per one successful drug candidate. [97]. Animal models are often used for these tasks but they have a significant disadvantage in the development of new drugs since their limited congruence with many human diseases severely affects their translational reliability [98].

There is thus a great need to develop new approaches to drug development. Multi-omics human data from deeply-phenotyped cohorts is one such direction and is considered one of the most promising potentials of analysing big Omics data in medicine [99]. First, analysis of large-scale health data may unravel new, unknown associations [100] and therefore may allow the discovery of new biomarkers and novel drug targets, for example, by mapping of existing genetic association findings to drug targets and compounds [101]. Second, it may be used to further evaluate the chances of success of drugs discovered and tested on animal models prior to the costly and timely stages of preclinical and clinical trials. Third, potential therapeutic interventions discovered by human data analysis with an established safety profile, such as nutritional modification or supplements and drugs with an existing Food and Drug Administration (FDA) approval, may be considered for a direct evaluation in human clinical trials (Figure 4). Finally, since most drugs affect only a subset of the treated target patient population, using human data to distinguish between responders and non-responders and prioritize responders to clinical trials can have great utility. Analysis of large scale human omics data has therefore the potential to accelerate drug development and reduce its cost. Indeed, it was previously estimated that selecting targets with evidence from human genetics data may double the success rate in the clinical development of drugs [102].

Systematic analysis of large-scale data by various computational approaches can also be used to obtain meaningful interpretations for repurposing of existing drugs [103]. For example, clinical information from over 13 years of EHRs originating 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 (ALS) [104]. Another example is using publicly available molecular data for the discovery of new candidate therapies for Inflammatory bowel disease (IBD) [105].

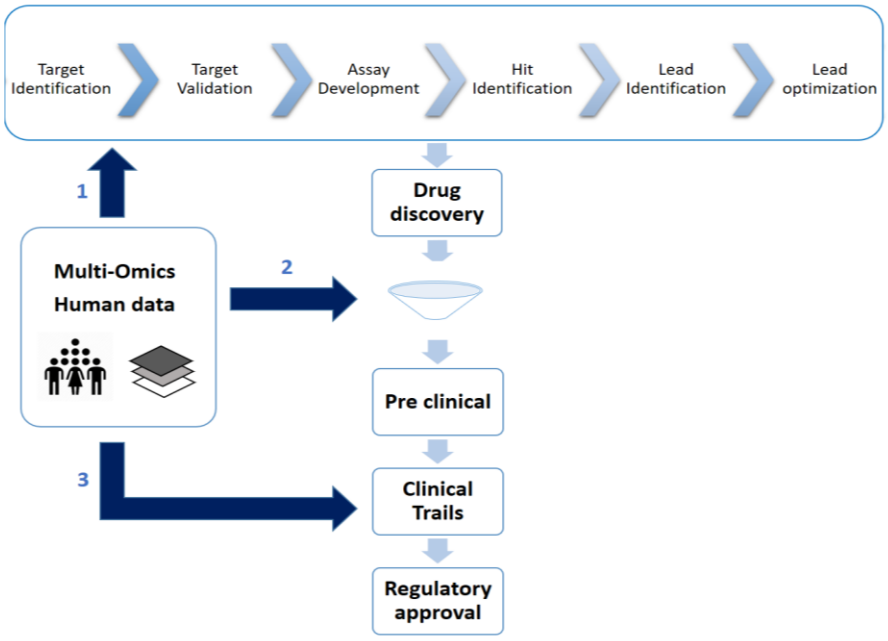

Figure 4: Using human-based omics data in drug development. Utilization of large scale human multi-omics data in the process of drug development may aid in: (1) Identification of new drug targets; (2) evaluation of drug candidates which were identified by animal models using humans data prior to preclinical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

5. *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, cut redundant tests [106], and can provide guidance for better health resources distribution [107]. 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.

6. *Disease phenotyping*

Novel phenotyping of disease and health and the study of variation between individuals is another potential of studying rich and novel types of data. For example, we previously characterized the variation between healthy individuals in response to food, based on deeply phenotyping a 1000-person cohort that included the first large-scale continuous glucose monitoring and gut microbiota profiling of healthy individuals [49].

Another potential is to refine current phenotyping of disease. For example, there have been recent attempts to refine the classification of type 2 diabetes and find sub-groups from available data [90,108]. Another example is Parkinson's disease (PD), where recent advances in genetics, imaging, and pathologic findings coupled with observed clinical variability, have profoundly changed the understanding of the disease. PD is now considered to be a syndrome rather than a single entity and the International Parkinson and Movement Disorders Society (MDS) have commissioned a task force for the redefinition of PD [109–111].

7. 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 [112]. However, caution should be taken, with careful assessments of how much of the change observed in the phenotype tested is actually due to variability within individuals [113]. It is not obvious that many of the medical questions of interest will be answered through big datasets. Historically, well designed and controlled (small) experiments were the primary drivers of medical knowledge, and the burden of showing a change in this paradigm is on the new methodologies.

Conclusion

Big data in medicine may give us the opportunity to view human health holistically, through a variety of lenses, each presenting opportunities to study different scientific questions. Here, we characterized health data by several different axes. The potentially scientific value of collecting large amounts of health data on human cohorts has recently been recognized and acted upon by different stakeholders, with a rapid rise in the creation of large scale cohorts aiming to maximize these axes and achieving various goals of understanding human health. Since maximizing each axis requires both resources and effort, it is inevitable that some axes come at the expense of others. Since delicate disease patterns may only be detected when the data includes a deep enough phenotyping (D) of a sufficient sample size (N), we believe that an interesting operating point is that of ‘deep cohorts’, whereby a medium sized cohort (e.g., tens of thousands) is profiled with a vast array of modern physiological and omics technologies.

Analysis of big data in health has many challenges and is in some sense a double-edged sword. On the 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 (“*Seeing the whole elephant*” analogy). In addition, real life evidence that it will translate to an improved quality of care which will benefit patients and the implementation of such models in current health systems is currently lacking. However, the potential to improve health care in many aspects 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 regarding disease pathogenesis, classification, diagnosis, prevention, treatment, and prognosis can be studied and potentially lead to new discoveries that may evantually revoulunaze medical practice.

References

1. CONSTITUTION OF THE WORLD HEALTH ORGANIZATION1.
2. Burton-Jeangros, C., Cullati, S., Sacker, A. & Blane, D. in *A Life Course Perspective on Health Trajectories and Transitions* (eds. Burton-Jeangros, C., Cullati, S., Sacker, A. & Blane, D.) (Springer, 2015). doi: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,** (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. Gould, A. L. Planning and revising the sample size for a trial. *Stat. Med.* **14,** 1039–51; discussion 1053 (1995).
10. 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).
11. Mason, C. E., Porter, S. G. & Smith, T. M. Characterizing multi-omic data in systems biology. *Adv. Exp. Med. Biol.* **799,** 15–38 (2014).
12. Cho, I. & Blaser, M. J. The human microbiome: at the interface of health and disease. *Nat. Rev. Genet.* **13,** 260–270 (2012).
13. Palsson, B. & Zengler, K. The challenges of integrating multi-omic data sets. *Nat. Chem. Biol.* **6,** 787–789 (2010).
14. Check Hayden, E. Is the $1,000 genome for real? *Nature* (2014). doi:10.1038/nature.2014.14530
15. 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).

16. Barker, D. J. In utero programming of chronic disease. *Clin. Sci.* **95,** 115–128 (1998).
17. Topol, E. J. Individualized medicine from prewomb to tomb. *Cell* **157,** 241–253 (2014).
18. Qiu, X. *et al.* The born in guangzhou cohort study (BIGCS). *Eur. J. Epidemiol.* **32,** 337–346 (2017).
19. Golding, Pembrey, Jones & The Alspac Study Team. ALSPAC-The Avon Longitudinal Study of Parents and Children. *Paediatr. Perinat. Epidemiol.* **15,** 74–87 (2001).
20. Howe, C. J., Cole, S. R., Lau, B., Napravnik, S. & Eron, J. J. Selection bias due to loss to follow up in cohort studies. *Epidemiology* **27,** 91–97 (2016).
21. Brieger, K. *et al.* Genes for good: engaging the public in genetics research via social media. *Am. J. Hum. Genet.* **105,** 65–77 (2019).
22. Kaprio, J. The Finnish Twin Cohort Study: an update. *Twin Res. Hum. Genet.* **16,** 157–162 (2013).
23. Magnus, P. *et al.* Cohort profile update: the norwegian mother and child cohort study (moba). *Int. J. Epidemiol.* **45,** 382–388 (2016).
24. Beesley, L. *et al.* The emerging landscape of epidemiological research based on biobanks linked to electronic health records: existing resources, analytic challenges and potential opportunities. (2018). doi:10.20944/preprints201809.0388.v1
25. Lau, B., Gange, S. J. & Moore, R. D. Interval and clinical cohort studies: epidemiological issues. *AIDS Res. Hum. Retroviruses* **23,** 769–776 (2007).
26. Chen, M. S., 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 Suppl 7,** 1091–1096 (2014).
27. 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).
28. Colditz, G. A., Manson, J. E. & Hankinson, S. E. The Nurses' Health Study: 20-Year Contribution to the Understanding of Health Among Women. *J Womens Health (Larchmt)* **6,** 49–62 (1997).
29. 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).

30. All of Us Research Program Investigators *et al.* The "All of Us" Research Program. *N. Engl. J. Med.* **381,** 668–676 (2019).

31. Hripcsak, G. *et al.* Observational health data sciences and informatics (OHDSI): opportunities for observational researchers. *Stud. Health Technol. Inform.* **216,** 574–578 (2015).

32. Rajkomar, A. *et al.* Scalable and accurate deep learning with electronic health records. *npj Digital Med.* **1,** 18 (2018).

33. 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).

34. Health linkage | UK Biobank. at <https://www.ukbiobank.ac.uk/health-linkage/>

35. 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).

36. Weber, G. M., Mandl, K. D. & Kohane, I. S. Finding the missing link for big biomedical data. *JAMA* **311,** 2479–2480 (2014).

37. Evans, R. S. Electronic health records: then, now, and in the future. *Yearb. Med. Inform.* **Suppl 1,** S48-61 (2016).

38. Tiik, M. & Ross, P. Patient opportunities in the Estonian Electronic Health Record System. *Stud. Health Technol. Inform.* **156,** 171–177 (2010).

39. Montgomery, J. Data sharing and the idea of ownership. *New Bioeth.* **23,** 81–86 (2017).

40. Rodwin, M. A. The case for public ownership of patient data. *JAMA* **302,** 86–88 (2009).

41. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3,** 160018 (2016).

42. Hewitt, R. & Watson, P. Defining biobank. *Biopreserv. Biobank.* **11,** 309–315 (2013).

43. OECD Glossary of Statistical Terms - Biobank Definition. at <https://stats.oecd.org/glossary/detail.asp?ID=7220>

44. Kinkorová, J. Biobanks in the era of personalized medicine: objectives, challenges, and innovation: Overview. *EPMA J.* **7,** 4 (2015).

45. 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).

46. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562,** 203–209 (2018).
47. 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).
48. Tapia-Conyer, R. *et al.* Cohort profile: the Mexico City Prospective Study. *Int. J. Epidemiol.* **35,** 243–249 (2006).
49. Zeevi, D. *et al.* Personalized nutrition by prediction of glycemic responses. *Cell* **163,** 1079–1094 (2015).
50. Rothschild, D. *et al.* Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555,** 210–215 (2018).
51. Zeevi, D. *et al.* Structural variation in the gut microbiome associates with host health. *Nature* **568,** 43–48 (2019).
52. 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).
53. 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).
54. Centre, B. & enquiriesukbiobank.ac.uk, S. UK Biobank: Protocol for a large-scale prospective epidemiological resource.
55. Cohen, I. G. & Mello, M. M. Big data, big tech, and protecting patient privacy. *JAMA* (2019). doi:10.1001/jama.2019.11365
56. Price, W. N. & Cohen, I. G. Privacy in the age of medical big data. *Nat. Med.* **25,** 37–43 (2019).
57. Tutton, R., Kaye, J. & Hoeyer, K. Governing UK Biobank: the importance of ensuring public trust. *Trends Biotechnol.* **22,** 284–285 (2004).
58. 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).
59. 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).

60. Shmueli, G. To Explain or to Predict? *Stat. Sci.* **25,** 289–310 (2010).

61. Geserick, M. *et al.* Acceleration of BMI in early childhood and risk of sustained obesity. *N. Engl. J. Med.* **379,** 1303–1312 (2018).

62. Obermeyer, Z., Samra, J. K. & Mullainathan, S. Individual differences in normal body temperature: longitudinal big data analysis of patient records. *BMJ* **359,** j5468 (2017).

63. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521,** 436–444 (2015).

64. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60,** 84–90 (2012).

65. Goldberger, A. L. *et al.* PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* **101,** E215-20 (2000).

66. Johnson, A. E. W. *et al.* MIMIC-III, a freely accessible critical care database. *Sci. Data* **3,** 160035 (2016).

67. Wang, S. *et al.* MIMIC-Extract: A Data Extraction, Preprocessing, and Representation Pipeline for MIMIC-III. *arXiv* (2019).

68. Ching, T. *et al.* Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **15,** (2018).

69. Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L. & Ranganath, R. Opportunities in Machine Learning for Healthcare. *arXiv* (2018).

70. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25,** 44–56 (2019).

71. Rajkomar, A., Dean, J. & Kohane, I. Machine learning in medicine. *N. Engl. J. Med.* **380,** 1347–1358 (2019).

72. Bengio, Y., Courville, A. & Vincent, P. Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35,** 1798–1828 (2013).

73. Weng, W.-H. & Szolovits, P. Representation Learning for Electronic Health Records. *arXiv* (2019).

74. Agniel, D., Kohane, I. S. & Weber, G. M. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ* **361,** k1479 (2018).

75. 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).

76. Pearl & Judea. *Causality*. (Cambridge University Press, 2009).

77. Johansson, F., Shalit, U. & Sontag, D. Learning Representations for Counterfactual Inference. (2016).

78. Smith, G. D. & Ebrahim, S. “Mendelian randomization”: can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32,** 1–22 (2003).

79. Hu, P., Jiao, R., Jin, L. & Xiong, M. Application of causal inference to genomic analysis: advances in methodology. *Front. Genet.* **9,** 238 (2018).

80. 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* (2019). doi:10.1016/S0140-6736(19)32008-2

81. Tomašev, N. *et al.* A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* **572,** 116–119 (2019).

82. 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).

83. Calvert, J. S. *et al.* A computational approach to early sepsis detection. *Comput. Biol. Med.* **74,** 69–73 (2016).

84. 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).

85. Avati, A. *et al.* Improving palliative care with deep learning. in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 311–316 (IEEE, 2017). doi:10.1109/BIBM.2017.8217669

86. Lipton, Z. C. The mythos of model interpretability. *Commun. ACM* **61,** 36–43 (2018).

87. Vogt, H., Green, S., Ekstrøm, C. T. & Brodersen, J. How precision medicine and screening with big data could increase overdiagnosis. *BMJ* **366,** l5270 (2019).

88. American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diabetes Care* **36 Suppl 1,** S67-74 (2013).

89. WHO | Obesity. at <https://www.who.int/topics/obesity/en/>

90. 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).

91. Young, A. I., Benonisdottir, S., Przeworski, M. & Kong, A. Deconstructing the sources of genotype-phenotype associations in humans. *Science* **365,** 1396–1400 (2019).

92. 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).

93. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538,** 161–164 (2016).

94. Kohane, I. S. Using electronic health records to drive discovery in disease genomics. *Nat. Rev. Genet.* **12,** 417–428 (2011).

95. Phelan, M., Bhavsar, N. A. & Goldstein, B. A. Illustrating Informed Presence Bias in Electronic Health Records Data: How Patient Interactions with a Health System Can Impact Inference. *EGEMS (Wash. DC)* **5,** 22 (2017).

96. What is the Phenotype KnowledgeBase? | PheKB. at <https://www.phekb.org/>

97. Brodniewicz, T. & Grynkiewicz, G. Preclinical drug development. *Acta Pol. Pharm.* **67,** 578–585 (2010).

98. Breyer, M. D. Improving productivity of modern-day drug discovery. *Expert Opin. Drug Discov.* **9,** 115–118 (2014).

99. Matthews, H., Hanison, J. & Nirmalan, N. "Omics"-Informed Drug and Biomarker Discovery: Opportunities, Challenges and Future Perspectives. *Proteomes* **4,** (2016).

100. Reshef, D. N. *et al.* Detecting novel associations in large data sets. *Science* **334,** 1518–1524 (2011).

101. Finan, C. *et al.* The druggable genome and support for target identification and validation in drug development. *Sci. Transl. Med.* **9,** (2017).

102. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47,** 856–860 (2015).

103. Pushpakom, S. *et al.* Drug repurposing: progress, challenges and recommendations. *Nat. Rev. Drug Discov.* **18,** 41–58 (2019).

104. Paik, H. *et al.* Repurpose terbutaline sulfate for amyotrophic lateral sclerosis using electronic medical records. *Sci. Rep.* **5,** 8580 (2015).

105. Dudley, J. T. *et al.* Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Sci. Transl. Med.* **3,** 96ra76 (2011).

106. Xu, S. *et al.* Prevalence and Predictability of Low-Yield Inpatient Laboratory Diagnostic Tests. *JAMA Netw. Open* **2,** e1910967 (2019).

107. Einav, L., Finkelstein, A., Mullainathan, S. & Obermeyer, Z. Predictive modeling of U.S. health care

spending in late life. *Science* **360,** 1462–1465 (2018).

108. 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).

109. Thenganatt, M. A. & Jankovic, J. Parkinson disease subtypes. *JAMA Neurol.* **71,** 499–504 (2014).

110. Lawton, M. *et al.* Developing and validating Parkinson's disease subtypes and their motor and cognitive progression. *J. Neurol. Neurosurg. Psychiatr.* **89,** 1279–1287 (2018).

111. 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).

112. Collins, F. S. & Varmus, H. A new initiative on precision medicine. *N. Engl. J. Med.* **372,** 793–795 (2015).

113. Senn, S. Statistical pitfalls of personalized medicine. *Nature* **563,** 619–621 (2018).