Prosperi et al. BMC Medical Informatics and Decision Making (2018) 18:139
https://doi.org/10.1186/s12911-018-0719-2

BMC Medical Informatics and Decision Making

DEBATE Open Access

Big data hurdles in precision medicine and precision public health

Mattia Prosperi[1*], Jae S. Min[1], Jiang Bian[2] and François Modave[3]

Abstract

Background: Nowadays, trendy research in biomedical sciences juxtaposes the term 'precision' to medicine and public health with companion words like big data, data science, and deep learning. Technological advancements permit the collection and merging of large heterogeneous datasets from different sources, from genome sequences to social media posts or from electronic health records to wearables. Additionally, complex algorithms supported by high-performance computing allow one to transform these large datasets into knowledge. Despite such progress, many barriers still exist against achieving precision medicine and precision public health interventions for the benefit of the individual and the population.

Main body: The present work focuses on analyzing both the technical and societal hurdles related to the development of prediction models of health risks, diagnoses and outcomes from integrated biomedical databases. Methodological challenges that need to be addressed include improving semantics of study designs: medical record data are inherently biased, and even the most advanced deep learning's denoising autoencoders cannot overcome the bias if not handled a priori by design. Societal challenges to face include evaluation of ethically actionable risk factors at the individual and population level; for instance, usage of gender, race, or ethnicity as risk modifiers, not as biological variables, could be replaced by modifiable environmental proxies such as lifestyle and dietary habits, household income, or access to educational resources.

Conclusions: Data science for precision medicine and public health warrants an informatics-oriented formalization of the study design and interoperability throughout all levels of the knowledge inference process, from the research semantics, to model development, and ultimately to implementation.

Background

The United States White House initiative on precision medicine stated that its mission is *"to enable a new era of medicine through research, technology, and policies that empower patients, researchers, and providers to work together toward development of individualized care"* [1]. Our ability to store data now largely surpasses our ability to effectively and efficiently learn from them and to develop actionable knowledge that leads to improvements in health outcomes. Precision medicine sprouts from big data and is the manifest evidence of such a dramatic change in scientific thinking. However, from its inception, precision medicine has been beleaguered with technical and sociopolitical challenges [2].

What is precision medicine?

The National Institutes of Health (NIH) defines precision medicine as the *"approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person"* [3]. The emphasis is placed on tailored prevention, diagnosis and treatment for each individual based on genetics, epigenetics, and other lifestyle considerations. The terms 'personalized,' 'stratified' and 'individualized' medicine have been often used interchangeably, but superseded lately by 'precision' [4]. Precision has been preferred *"to emphasize the new aspects of this field, which is being driven by new diagnostics and therapeutics"* [5]. Nonetheless, the debate on terms and definitions is still open [6].

A classic example of precision medicine is the customization of disease treatment for a single individual. In the old paradigm of one-size-fits-all medicine, an effective treatment is the treatment known to benefit

* Correspondence: m.prosperi@ufl.edu
[1]Department of Epidemiology, College of Medicine & College of Public Health and Health Professions, University of Florida, Gainesville, FL 32610, USA
Full list of author information is available at the end of the article

most of the target population, which is usually captured using the notion of *number needed to treat*. The number needed to treat (NNT) is a measure indicating the average number of people who need to be treated to avert one additional bad outcome. For instance, a commonly used treatment for cholesterol has a NNT of 20, which means 1 out of the 20 who are treated will actually yield benefit from the said treatment [7]. The rest of the population will not benefit from the treatment, and may even incur adverse effects. This exemplifies the need for customized treatment based on variables such as genetics, ethnicity or lifestyle. The underlying assumption is that precision medicine will provide tailored health care to patients and will yield lower rates of associated adverse outcomes. Although precision medicine aims at prevention, diagnosis and treatment, the main efforts have been centered around precision pharmacogenomics and the delivery of drugs based on patients' specific genetic markers. For instance, the administration of drugs like clopidogrel is based on an individual's genetic susceptibility for speedier metabolism [8] or risk for hypersensitivity to antiretroviral therapy abacavir is calculated based on a genetic test [9]. In the precision medicine paradigm, given detailed patient characteristics, it is possible to more accurately predict the expected effect of each treatment option and, thus, to optimize care.

Are clinical trials precision medicine?

One may argue that clinical trials have always been operating with a precision medicine paradigm, by testing therapies on homogeneous set of patients who are most likely to benefit and yield the most favorable outcomes from the drug. However, even with randomization, participation in clinical trials is not uniform across demographic, social, genetic –excusing Mendelian randomization [10]– and other factors that influence health. Historically, women, minorities, children and pregnant women have many times been excluded or underrepresented in clinical trials, and although this landscape is changing, it has not yet reached the levels of representativeness of the general population [11, 12]. Therefore, clinical trials have been 'precise' only for a subset of the population. Additionally, the costs associated with sufficiently-powered clinical trials stratified across all possible outcome modifiers make them prohibitive as a cost-effective precision strategy.

More variables, more observations

In order to be precise, medicine must revolve around data, especially in generating, linking, and learning from a variety of sources. This means going beyond genetics and exploring data that may not be traditionally thought of as being related to health and disease. However, resources need to be included as a key variable of precision medicine, regardless of the health system one considers. Indeed, health monitoring can quickly become expensive, and thus, cost-effective strategies need to be identified across the continuum of care. For instance, while there are biomarkers that are static in nature (e.g. a specific genetic variant), others change over time and need to be evaluated periodically. In an ideal world where a plethora of markers can be used to predict future health status with high precision, a cost-effective set should be identified in order to guarantee the same performance with minimal burden.

The main objective of this paper is to 1) explore the evolution of medicine and public health in a data-rich world; 2) present the current main hurdles to have precision health deliver on its promises; and subsequently 3) propose a modeling framework to remove some of these barriers to precision medicine and precision public health, improving health outcomes and reducing health disparities. To help the reader throughout the sections, we have prepared a summary list of the arguments of our debate, listed in Table 1.

The evolution of precision medicine

Redefining precision care

Individualized treatment, e.g. tailored pharmacotherapy, is not the sole component of precision medicine. From a utilitarian point of view, it may be useful to break down precision medicine by its components across the continuum of care (Fig. 1a), to be met under specific time constraints:

- Disease prevention, or prediction of disease risk before the disease symptoms manifest,
- Differential diagnosis, or timely/instantaneous identification of an illness, and
- Disease treatment, i.e. strategies to cure or optimally treat once disease has been identified.

These components reflect the move from a focus on treatment only in health care to include also prevention, as well as prognosis, and post-disease survivorship, as critical aspects of medicine and health in general.

In the context of disease prevention, the goal of precision medicine is accurate prediction of disease risk – even years in advance. For instance, many risk factors associated with lung cancer are well understood, and even though behavior change is difficult, the timeframe is sufficient to intervene in order to decrease risks. For all diseases with potentially severe health outcomes, whose etiology is not entirely understood, prediction modelling should be used to identify disease markers as early as possible. However, prediction models are only useful if they can include risk factors that are modifiable –such as dietary habits and lifestyle (because genes, age, race are not). Such models would allow changing the

Table 1 Hurdles in precision medicine and precision public health within data, study, model development, and deployment phases

Precision medicine

- Concentration on individualized treatment and neglect of time component of predictions, i.e. early risk vs. differential diagnosis vs. post-treatment survival
- Too much focus on genetics and –omics
- Research on actionable factors vs. immutable risk factors
- Integration of multi-omics
- Integration of multi-domain data (e.g. genetics, diet, lifestyle, social)

Precision public health

- Definition of target units (e.g. ethnic groups, geographic zones, social groups)
- Conflict with precision medicine, i.e. individual-centric objectives (benefit of the single may not translate into benefit of the population)
- Population-level outcomes

Data sources	Study designs	Prediction modelling	Translational relevance
• Heterogeneous data sources • Unstructured data sources • Lack of data on social determinants of health • Measurement issues (e.g. incompleteness, inaccuracy, imprecision in self-reported data) • Privacy and security • Cost • Limited adoption of common data models	• Semantic data integration (i.e. linking data elements by their meaning) • Large longitudinal cohorts • Ontology integration • Ontology appropriateness (e.g. ontologies made for billing vs. for diagnostic purposes) • Semantic interoperability • Automated study design	• Biases of all sorts (e.g. protopathic) • Confounding • Causal inference • Black-boxes vs. white-boxes (i.e. interpretability vs. performance) • Complexity-based model selection • Benchmark development • Pragmatic interoperability (reproducibility, replicability, generalizability)	• Limited individual empowerment • Disconnect from relevant clinical research • Personal health record/health avatar (besides provider's electronic records) • Acceptance of artificial intelligence as integral part of doctors' tools • Learning systems • Ethical usage and dissemination of modelling algorithms • Redefining disease phenotype

odds of the disease onset, and possibly with enough time for an intervention. In many modelling approaches, as we will see in the next sections, the value of the data from the modifiability point of view is not taken in the account, as well as from a 'comprehensibility' point of view (i.e. understanding the mechanics of the underlying biological processes by decomposing the model functions).

With differential diagnosis, the timeframe is reduced to a matter of days or even hours. Acute abdominal pain can have very different etiology, ranging from aortic aneurysm to kidney stones, or peritonitis. Mistaking a chronic condition, e.g. Meniere's, can severely affect quality of life [13]. Besides genetic markers, in this case we can think of high-sensitivity tests with rapid

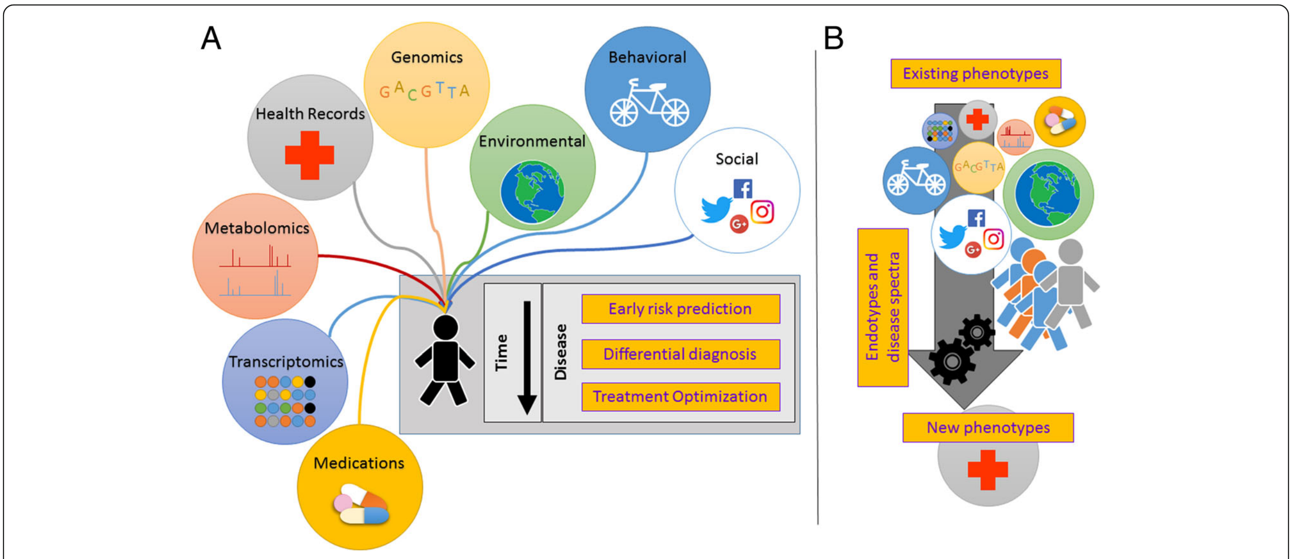


Fig. 1 Revisitation of precision medicine. **a** Precision medicine moves from a genetics-centered personalization of treatment on to a dynamic risk assessment and optimization of current and future health status through immutable (e.g. genetics) and actionable factors (e.g. behavior). **b** Disease phenotypes are reclassified on the basis of new system-level evidence, identifying pathophysiological endotypes associated with common, known phenotypes. The logos are trademarks of their respective companies and institutions, and their uses do not represent affiliation or endorsement

turnaround time, like metagenomics sequencing to screen for multiple pathogens. Many diseases and diagnoses are still not clearly defined: *"Descriptions of disease phenotypes often fail to capture the diverse manifestations of common diseases or to define subclasses of those diseases that predict the outcome or response to treatment. Phenotype descriptions are typically sloppy or imprecise"* [14, 15]. For instance, asthma is an umbrella disease, with possibly different underlying endotypes [16]. Rheumatoid arthritis and its related gamut of symptoms, or other types of autoimmune conditions, as well as Alzheimer's disease are considered system-level illnesses [17]. Therefore, one additional constituent of precision medicine is the notion of 'precise phenotyping' (Fig. 1b).

The last point, i.e. disease treatment, is the last stand against unfavorable health outcomes. It necessarily builds upon the former two and moves forward by adding more complexity, i.e. the space of treatments. Different outcomes can be obtained by running prediction models that set a patient's status (e.g. genes, metabolic profile, drug exposures) and vary treatments (e.g. new drugs, diet, physical activity). Treatment optimization, seen as an operational research problem, explores this outcome prediction space looking for the most favorable ones.

Genetic epidemiology: the big short

The widespread availability of sequencing methods along with a drastic reduction of their associated costs were largely responsible for the rise and evolution of precision medicine. Today, genome-wide sequencing can cost about $1000, down from close to $98 M in 2001. Moreover, several commercial companies are offering services that provide partial genome sequencing for a little over $100, along with mapping to demographic traits and specific disease conditions (yet with unproven clinical utility), which isn't without raising concerns related to privacy of health information [18]. Although a genome stays relatively immutable during a lifetime, a genomic screening obtained at birth or before birth will be optimal for the most accurate disease prediction [19]. Despite the initial enthusiasm for genetics-focused precision medicine, the results have been underwhelming and have not delivered on its promises so far. The predictive ability and ensuing clinical utility of risk assessment from genetic variations has been found to be modest for many diseases [20, 21], and genome-wide association studies (GWAS) have not led to understanding genetic mechanisms underlying the development of many diseases [22].

Among the shortcomings of GWAS, one is the *missing heritability* problem –heritability is a measure of the proportion of phenotypic variation between people explained by genetic variation– for which single genetic variations cannot account for much of the heritability of diseases, behaviors, and other phenotypes.

Another limitation of GWAS relates to studying a single phenotype or outcome (often imprecise, as we pointed out in the previous section), and accounting for heterogeneous phenotypes would require studies massive in size [23]. Other GWAS issues include design, power, failure of replication, and statistical limitations. In practice, only univariate and multivariable linear regression is performed. Looking at gene-gene interactions, and including other variables rather than basic demographics or clinical traits is rarely done and often computationally burdensome.

There are very few examples of high-effects common genetic variants influencing highly-prevalent diseases, and common genetic variants usually have low predictive ability. The rarer a genetic variant, the harder it is to power a study and ascertain the effect size. There are rare high-effect alleles causing Mendelian diseases and a glut of low-frequency variants with intermediate effects. Low-effect rare variants are very difficult to find and may be clinically irrelevant, unless implicated in more complex pathways [24]. In fact, it is known that gene expression pairs can jointly correlate with a disease phenotype, and higher-order interactions likely play a role too [25–27]. Few algorithms have been proposed to seek jointly-expressed genes, and existing methods are computationally inefficient [28].

However, these issues are only partially responsible for precision medicine not yet meeting its original promises. Indeed, for precision medicine and precision public health models to be valid and effective, incorporating and testing factors beyond genetics is key. While genetics remains mostly static over time, other health-related factors are constantly changing and need to be evaluated periodically. Epigenetics, e.g. methylation data, which has a time component, can contribute to a relevant portion of unexplained heritability [29]. Cheaper and faster production of sequence data with next-generation sequencing technologies has opened the post-GWAS era, allowing for a whole new world of –omics [30, 31].

Domain-wide association studies

The GWAS revolution, and arguably saturation, has brought a surfeit of epigenome-wide –methylation-wide, transcriptome-wide– [32], microbiome-wide [33], and environment-wide association studies [34]. Interestingly, phenome-wide association studies reverse the canon, as all health conditions found in medical histories are used as variables and associated to single genetic traits [35].

Genomics, transcriptomics, metabolomics, and all other –omics can be seen as input *domains* to a prediction model. Merging of two or more domain-wide association studies is the next step toward a better characterization of disease mechanics and risks [36].

However, modelling and computational challenges arise with multi-domain integration, because of increased dimensions, variable heterogeneity, confounding, and causality. Formalizations of cross-domain-wide association studies, under the general umbrella term of *multiomics*, have been proposed [37–39]. Despite the cheaper and faster production of sequence data, most of the multiomics studies are limited by small samples: in general, the more heterogeneous the experimental data to be generated or the data sources to be included in the study are, the more difficult it is to get larger sample size.

The most interesting utility of multiomics, rather than prediction of health outcomes, is their 'unsupervised' analysis, i.e. the identification of patterns/endotypes that can help unveiling biological pathways, and eventually redefine disease spectra and phenotypes. However, there is mounting evidence that to ensure that precision medicine and precision public health deliver on their promises across the care continuum, we need to go beyond the –omics.

Beyond traditional domains

In the era of precision medicine, multi-domain studies need to extend beyond 'omics' data and consider other domains in a person's life. Specifically, genetic, behavioral, social, environmental, and clinical domains of life are thought to be the five domains that influence health [40]. Further, the ubiquitous nature of Internet access and the widespread availability and use of smartphone technologies suggest that the clinical domain can be significantly enhanced with patient-generated data, such as physical activity data, dietary intake, blood glucose, blood pressure, and other similar variables that can be seamlessly collected using smartphones and wearable devices [41].

Moreover, such tools, combined with social networks platforms provide a window into the behavioral and social domains of health, data-rich environments that need to be considered in the context of precision medicine and precision public health, to create a 'digital phenotype' of disease [42]. For instance, images from Instagram have been used to ascertain dietary habits [43] in lieu of a food diary or dietary intake questionnaires, which can be inaccurate, and are cumbersome and time-consuming; Instagram, again, has been used to identify predictive markers of depression [44]. The passively collected data from Twitter can be used for insomnia types characterization and prediction [42]. Research into the environmental domain has shown that the environment in which we live in impacts our health and mortality [45]. However, research using non-traditional health-related data from these domains have been conducted with some success as well as with some controversy [46, 47].

The health avatar

As precision medicine fundamentally reduces to fine-grained, individual-centric data mining, the observational unit of such data, and pivot for domain linkage, can be defined with the theoretical model of the *health avatar*. The health avatar is a virtual representation of a person with all their associated health information (Fig. 2), and intelligent ways to manage and predict their future health status [48].

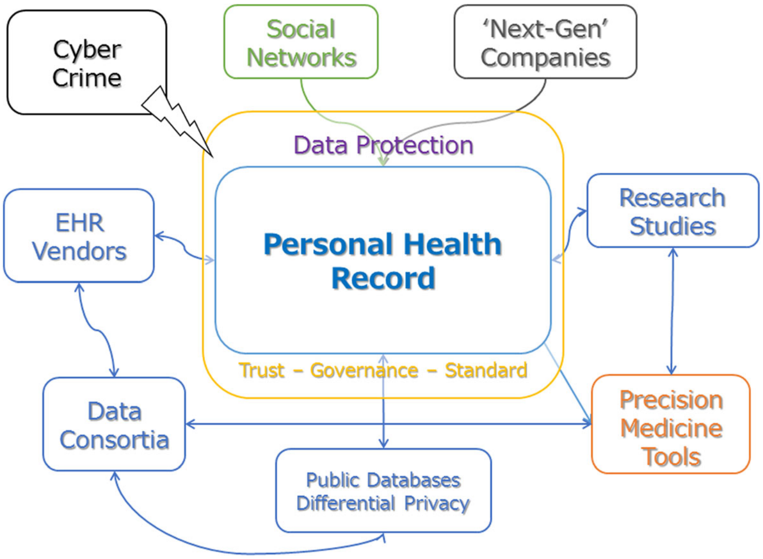


Fig. 2 The health avatar: a virtual representation of a person with all their associated health information, and intelligent ways to manage and predict their future health status. The health avatar is centered on the personal health records and integrated with healthcare, commercial governance, and research entities

Even with the widespread use of electronic health records (EHR) and integrated data repositories, individuals are generally detached from their health information and opportunities to be actively involved in research remain limited, despite initiatives such as Apple's HealthKit. Well-known barriers to linking and efficiently exploiting health information across different sites slow down healthcare research and the development of individualized care. Further, EHR are not translationally integrated with diagnostic or treatment optimization tools. A doctor can get and transfer lab results online, but then diagnoses are often made in a traditional manner, based on average population data.

The personal health record (PHR) is a collation of all health information from different healthcare providers or other sources that is stored in the cloud, a directly accessible property of the individual [49, 50]. The PHR is complementary to the EHR, which is usually stored at the provider level, with vendors' software, such as Epic [51] or Cerner [52]. However, the health avatar should not simplistically be identified with the PHR, as the PHR is inherently passive, with little involvement from the patient. We now propose a model of what the modern health avatar should be, in an era of large patient-generated data sets. An individual can see their health information using a provider's PHR, but cannot easily merge the information with data from other providers nor ask a provider to upload their data from the EHR to the PHR simply during a doctor's visit, e.g. via a smartphone app. An intelligent algorithm that matches people to research studies based on their full medical history does not exist yet. Both doctors and patients who are interested in computer-aided diagnosis, usually have to upload information to a third-party service (e.g. to analyze susceptibility to antibiotics). Finally, data shares are cumbersome, not only in terms of steps required to respect ethical principles, practice, and to protect human subjects, which are necessary but could be modernized, but also because the only data considered reliable are those coming from EHR. This means that big data shares happen solely at the population level via institutional or corporations' liaise. Research and analytics that follow are not streamlined; the long-awaited *research objects* –semantically rich aggregations of resources that bring together data, methods and people in scientific investigations– are still in their infancy [53]. Integration of different types and sources of data should retain original context and meaning while meaningfully mapping their relationships to other health-related variables; such semantic integration will need to be flexible and comprehensive.

Physical data integration of EHRs requires enormous efforts and resources, but currently is the most successful approach to health information linkage because it is supported by rigorous governance standards and solid infrastructure. Efforts like the national patient-centered clinical research network [54] is a prominent example. Data sharing for matching research participants, one of the long-awaited prerogatives of NIH, is finally being exploited, via ResearchMatch [55].

The health avatar should link all and new types of health-related data, from genomics, to the myriad of -omics, mobile and wearable technology-based, and environmental sources. These data capture information from other domains which impact health far greater than clinical care alone. Such integrations have already begun around the world with healthcare systems like Geisinger conducting genetic sequencing and returning some of the results to the patients and with initiatives such as electronic Medical Records and Genomics (eMERGE) and Implementing Genomics in Practice (IGNITE) networks [56]. However, these efforts have been limited to genomics. More generally, the health avatar should be able to connect with and exploit non-EHR information potentially useful for health assessment, even coming from highly unstructured sources, such as social media. This is exemplified recently with Epic partnering with Apple to allow Apple's HealthKit to display patient's EHR data. Epic's App Orchard also allows the collection of wearable technology data and storage into the EHR. For instance, an artificial intelligence tool could process images from Instagram and Facebook/Twitter posts to ascertain dietary habits, and this information can then be used to populate a food questionnaire, encoded into some type of structured information and stored in the EHR. Moving out of the individual level, environment-level information pertinent to the individual –for instance through residence ascertainment or mobile geolocation– could also populate EHR fields, storing information such as exposure to allergens and pollutants.

However, the collation of non-standard data, e.g. momentary ecological assessment via Twitter, Facebook, or smartphone GPS monitoring, is prone to serious privacy and security concerns. Ubiquitous approaches must be foreseen, as in the *Internet of things* [57, 58]. Data integration, and even more data share, must be secure to meet popular support. In this sense, the research in *differential privacy* aims at developing new algorithms not only to protect identities, but also to generate masked or synthetic data that can be shared publicly and freely used for preliminary research [59–61]. While differential privacy has facilitated data sharing, it remains challenging to safely anonymize data while preserving all their multivariate statistical properties [62]. The individual-centric approach of the health avatar can facilitate the match of individuals with research programs, with blurred boundaries between clinical care and research, while respecting ethics but modernizing informed consent concepts.

In terms of active features, i.e. not only data storage, the health avatar would feature linkage to personalized predictive tools for health status. Within the context of appropriate ethics bylaws and informed consents, health avatars could directly feed individual-level health information to multiple research projects for creating new and more accurate precision medicine tools. This would require data privacy and protection measures to avoid identity or data theft and misuse. Further, wide access to patient-generated data, along with integration with clinical and health databases provide a unique opportunity to expand precision medicine to the population level. We discuss this specific expansion in the following section.

Precision public health

The Director of Office of Public Health Genomics at the Centers for Diseases Control and Prevention (CDC) defined 'precision' in the context of public health as *"improving the ability to prevent disease, promote health, and reduce health disparities in populations by: 1) applying emerging methods and technologies for measuring disease, pathogens, exposures, behaviors, and susceptibility in populations; 2) developing policies and targeted implementation programs to improve health"* [63]. Top priorities included: early detection of outbreaks, modernizing surveillance, and targeted health interventions. To achieve such improvements, comprehensive and real-time data to learn from are necessary. Epidemiology must expand surveillance on to multiple, different information domains, such as the Internet and social media, e.g. infodemiology [64]. Big data does not only mean large sample size or fine-grained sampling, but also large variety of variables. So far, the big data emphasis is on sequencing genomes population-wide [65, 66], but research has started to consider other domains, such as in integrating classical surveillance with geospatial modelling [67].

In yet another step-by-step guide to precision public health –focused on developing countries– better surveillance data, better data analyses, and rapid actions are urged: again, big data is the key, with emphasis on public data sharing, and on the data attributes of speed-accuracy-equity. Notably, this is an epidemiological projection of the canonical big data characteristics, known as the Vs [68, 69].

Winston Churchill famously stated that *"healthy citizens are the greatest asset any country can have"*, and to achieve health for all citizens, there needs to be a transition from precision medicine, which is individualized, to precision public health. In fact, precision medicine can be used to improve an individual's health, but this does not necessarily translate into a uniform benefit for the population [70]. For instance, a precision medicine model tuned for majority of a population may improve the average health outcomes overall yet neglect minorities. To some extent, the term precision put next to population-wise priorities seems conflictual, and this may be due to application of single precision public health model to an entire population, rather than use of multiple segmented/cluster level models. Another fuzzy aspect of current precision public health approach is the lack of consensus on observational units used for inference or intervention [71]. Is it the individual? Is it a common geographic area? Is it a particular subpopulation? A theory-based approach in this sense would be useful, as we will show in the next section.

Precision public health has to face societal challenges, including racial disparities (both in terms of welfare and genetic background), environmental niches (e.g. tropical climate with higher rates of arboviral diseases, industrial areas with high pollution), and general ethical concerns (religious beliefs, political views). An individual-centric model such as the health avatar here poses a number of limitations, because it may lack of higher-level dynamics happening at the societal-environmental level (Fig. 3). Interestingly, such dynamics can also influence the individual itself, and therefore should be accounted for and projected on to the person-centric models.

Big data modelling for precision medicine and precision public health

Semantic integration

Barriers to linking and efficiently exploiting health information across different sites slow down healthcare research and the development of individualized care. Different EHR systems may independently define their own data structural formats. This independent and heterogeneous management poses challenges in information mapping and encoding, e.g. merging data from multiple EHR systems or from different standardization procedures without access to the original data.

Data integration across multiple domains and sources is a daunting task due to at least three factors: 1) the heterogeneity in the syntax of the data such as the different file formats and access protocols used, 2) multiple schema or data structures, and more importantly 3) the different or ambiguous semantics (e.g. meanings or interpretations). Substantial effort is required to link different sources due to lack of clear semantic definitions of variables, measures, and constructs, but it can be eased by *semantic interoperability,* which allows exchange of data with unambiguous, shared meaning [72–74].

A common approach in semantic data integration is through the use of *ontologies.* Building upon a standardized and controlled vocabulary for describing data elements and the relationships between the elements, an ontology can formally and computationally represents a domain of knowledge [75, 76]. With a universal conceptual representation of all the information across different sources, a semantic integration approach allows us to

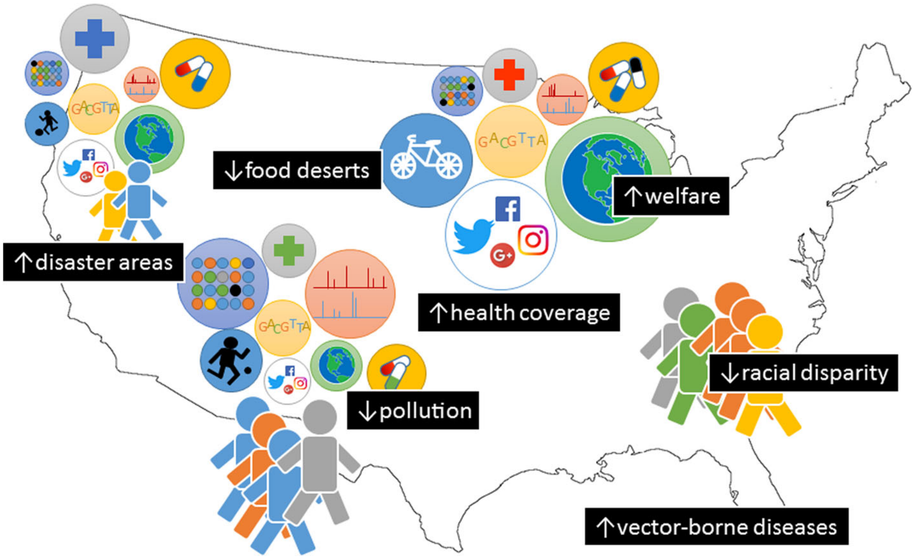


Fig. 3 Precision public health. Community, societal and ecological factors must be accounted on top of the individual-based, fine-grained approach for precision medicine. The map is an edited version of a Wikimedia Commons image (https://commons.wikimedia.org/wiki/File:United_States_Administrative_Divisions_Blank.png, licensed under the Creative Commons Attribution-Share Alike 3.0 Unported)

bridge the heterogeneity of data across multiple sources and domains. Many biomedical ontologies are already available and widely used in medicine, e.g. the International Classification of Diseases (ICD) or the Systematized Nomenclature of Medicine and Clinical Terms (SNOMED CT) [77, 78]. Nevertheless, an unified ontology-driven data integration framework is needed to accommodate the growing needs of linking and integrating data from multiple domains. Going beyond traditional approaches of using *common data elements* and *common data models (CDM)* [79], such as the international efforts in building the Observational Medical Outcomes Partnership (OMOP) CDM [80], an ontology-driven data integration framework can be used to represent metadata, create global concept maps, automate data quality checks, and support high-level semantic queries. Further, research on the semantics of EHR improves not only data integration and interoperability, but can also advance the science on disease phenotyping [81–84].

Moreover, ontologies can be used to facilitate a formal documentation of the data integration processes (e.g. through encoding the relationships between the variables to be integrated across different sources). Doing so can have significant impact on research rigor, transparency, and reproducibility among scientists as well as data reusability and flexibility.

Semantic integration can occur at different levels of healthcare research, not only at the data level with EHR. As mentioned, study designs on integrated data sources need to be supported by proper semantics. In Fig. 4 we summarize the semantic integration paradigm at different levels: (i) the data level, integrating both EHR and PHR data sources (inter-domain); (ii) the concept level, mapping terminologies and ontologies (domain--contextual); (iii) the study design level, enabling standard operating procedures and reproducibility on other sources (domain-contextual); (iv) the inference level, identifying proper statistical learning methods upon study design, scaling analyses on high-performance computing, and building up models and applications for the public health benefit (trans-domain). Semantic integration allows modularity (e.g. addition of new data or ontology components), flexibility (e.g. modification of existing study designs or execution in different environments), and transparency (e.g. reproducibility of results, validation, enhancement of models).

For instance, interoperable semantics and research objects have been the driver to the 'asthma e-lab' project [85]. As a secure web-based environment to support data integration, description and sharing, the e-lab is coupled with computational resources and a scientific social network to support collaborative research and knowledge transfer.

Another relevant example is the Observational Health Data Sciences and Informatics (OHDSI) [86] consortium, whose goal is *"to create and apply open-source data analytic solutions to a large network of health databases to improve human health and wellbeing."* OHDSI uses the OMOP common data model and features a suite of applications for streamlining integration of EHR, data quality assessment and cleaning (ACHILLES), standardized vocabulary for OMOP (ATHENA), data query and cohort identification (ATLAS), and analytics (CYCLOPS).

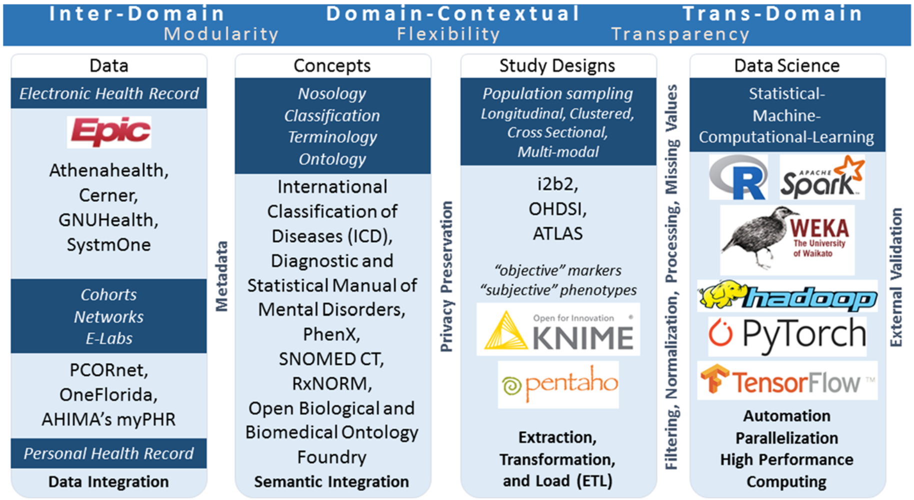


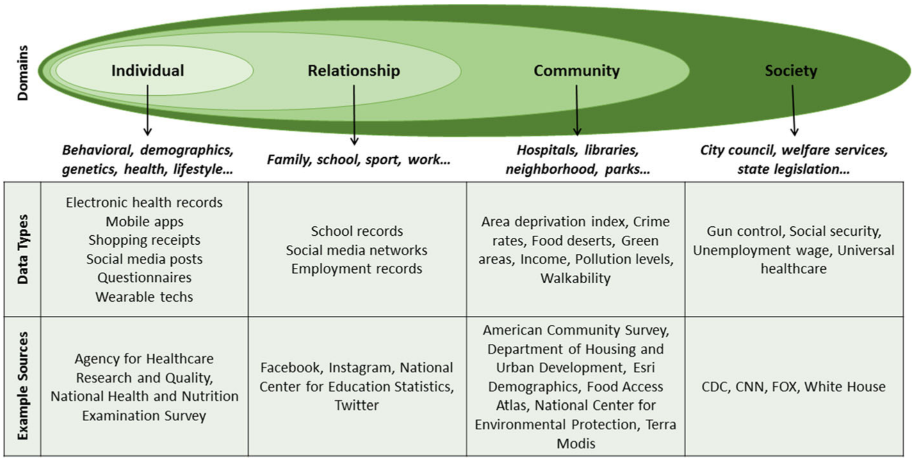
Fig. 4 Semantic integration on data, study design and inference. The logos are trademarks of their respective companies and institutions, and their uses do not represent affiliation or endorsement. TensorFlow, the TensorFlow logo and any related marks are trademarks of Google Inc. The logos are used for informative purposes only, and the list included here is not exhaustive

With semantic interoperability standing, we move on to study design theorization for precision medicine and precision public health.

Study designs: hollow learning, shallow design

With the advancements in technology and data linkage, single-domain research is being superseded by multi-level, multi-domain studies. Such increase in complexity and heterogeneity of studies affects also their design, in the case of both prospective and observational designs. Especially for observational studies, there is huge amount of data potentially available, but the access and use of such heterogeneous data sources must be rationalized to tackle bias, identify actionable inputs, and consider ethical needs.

In psychology research, it has been proposed that data-driven studies should be guided by an etiological theory in terms of study design [87]. These theories are grounded on evaluating scientific evidence as causal pathways of disease. Hybridization of using theory to guide design ('top-down' approach) with data-driven research ('bottom-up' approach) can be very useful for development of multi-level and multi-domain prediction models, encompassing individual and population levels. Several conceptual models exist that can be used, such as the social-ecological model or the multi-causality model [88, 89]. The challenge when using such models is to identify the sources of information for each component and to link the data, as we just discussed in the health avatar and semantic integration sections. In Fig. 5,



	Individual	Relationship	Community	Society
Domains	*Behavioral, demographics, genetics, health, lifestyle…*	*Family, school, sport, work…*	*Hospitals, libraries, neighborhood, parks…*	*City council, welfare services, state legislation…*
Data Types	Electronic health records Mobile apps Shopping receipts Social media posts Questionnaires Wearable techs	School records Social media networks Employment records	Area deprivation index, Crime rates, Food deserts, Green areas, Income, Pollution levels, Walkability	Gun control, Social security, Unemployment wage, Universal healthcare
Example Sources	Agency for Healthcare Research and Quality, National Health and Nutrition Examination Survey	Facebook, Instagram, National Center for Education Statistics, Twitter	American Community Survey, Department of Housing and Urban Development, Esri Demographics, Food Access Atlas, National Center for Environmental Protection, Terra Modis	CDC, CNN, FOX, White House

Fig. 5 The social-ecological model with associated information domains and data sources for a multi-domain study design

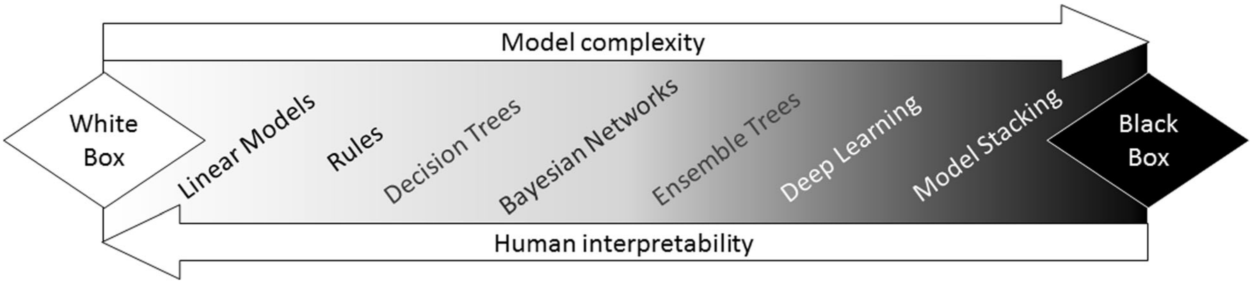


Fig. 6 Machine learning models: white- and black-boxes. Increasing model complexity can lead to better approximation of functions and enhance prediction performance, but can lead to a decrease in interpretability of the model

we show the social-ecological model, the information domains, and a number of data sources (mostly available in the United States, for illustrative purposes) that can be used to extract relevant attributes for the domain dimensions.

The advantage of using a theoretical model is that it is possible to deconstruct the prediction model to test hypotheses or identify new areas that need further investigation. For example, suppose we use the social-ecological model and integrate individual-level EHR and genetic markers with community-level social and ecological indicators, over a specific time horizon, to determine population risk of acute or chronic asthma. Certain variables in the individual- or the community-levels may be found to contribute to increased risk, and through cross-domain interactions, the percentage of variance explained may increase. Furthermore, variables in each domain can be examined to see if they are actionable or immutable (e.g. environmental exposures vs. genetics) and ethically usable or not (e.g. neighborhood deprivation score vs. racial profiling). This information can be exploited to determine a proper risk model and to select factors that can be modified to reduce the risk of disease.

One of the biggest hurdles in study design, especially for observational or retrospective studies, relates to effectively identifying and addressing bias. With big data, this issue is severe, because of data collection heterogeneity, source verification, and sampling bias among others. Researchers must be wary of the ‘big data hubris’ or *“that big data are a substitute for, rather than a supplement to, traditional data collection and analysis”* [46]. With EHR, bias overwhelms randomization. EHR data are inherently biased by the patient population structure, frequency of health care visits, diagnostic criteria, and care pathways; drug prescription records mostly reflect indication or protopathic bias. Even the most advanced statistical methods cannot disentangle bias, but they can learn it very precisely. Therefore, feeding a deep learning architecture with raw EHR data may be a very bad idea, although it yields amazing prediction performance [90–93]. In fact, *“biased algorithms are everywhere, and no one seems to care”* [94]. The problem is not novel and becomes dangerous if used for decision making [95]. Besides tragicomic revamping of phrenology through deep learning [96], ProPublica’s assessment of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm, a tool used to predict a person’s risk of recidivism, is a serious example of bias-learning models [97].

Prediction modelling: interpretability vs. performance

Another important challenge in use of big data for precision public health is the utility of inferred models, i.e. *“Do big data lead to big models?”* ‘Big’ models contain many variables and in nonlinear or highly complex ways, and such machine learning models can yield easily interpretable results or excellent prediction, but not necessarily both at the same time. In spite of the potentially higher accuracy in predicting disease diagnoses and health outcomes, many machine learning methods are usually regarded as non-transparent to the end user and labeled as *black-boxes.* In opposition, *white-boxes* are human-interpretable models, such as risk scores or diagnostic rules. Although black-box models may provide a very precise calculation of the probability of a target event or outcome, they are often regarded with skepticism due to the lack of consideration for causal pathways (Fig. 6). However, when integrated seamlessly in EHR as a clinical decisions support system and if they can identify clinically actionable features, they can be more acceptable [98].

Management of the tradeoff between interpretability and prediction performance is often neglected when developing frameworks for predictive analytics, but it can be critical for deploying the models in clinical practice [99]. One possible way to balance between white- and black-boxes is to use the more complex strategy known as the *super learning* framework [100], or *stacking*, and deconstruct its components. Essentially, the super learning approach fits and stacks many different models together on the data and selects the best weighted combination. Although super learning approaches are thought to have maximal prediction accuracy and minimal interpretability, deconstruction into digestible components is a necessary step for interpretability and thus, clinical utility. This can be extended to test various domains to include in the model, to optimize the model, and to guide future explorations of the data (Fig. 7).

For instance, the super ICU learning algorithm (SICULA) has been constructed for mortality prediction [101]. Post-hoc tools to identify the importance of individual variables can break down the complexity of black-box models like the random forest or neural networks [102, 103].

Model complexity is not universally defined, but indices like the Vapnik-Chervonenkis dimension can be used [104]. When selecting models on the basis of their complexity, there are two advantages: 1) performance thresholds can be set on the basis of clinical utility such that a more interpretable model that is less accurate than a more complex one could be chosen if it meets the required sensitivity or specificity; and 2) model simplification and interpretation can help in understanding findings to develop new etiological or mechanistic hypotheses. Nonetheless, the picture is not as simple: there is no guarantee that the combined information induced by a super learner will be straightforward to deconstruct; if the final models are deep neural networks, they will still be very challenging to interpret. The interpretability of complex and/or stacked models will still be limited by the inherent interpretability of the underlying components and functions. Downstream analysis like variable importance ranking or partial dependence plots may be helpful, but these solutions are highly model-dependent and can be biased by numerous factors (such as variable collinearity).

Modelling interoperability

Besides semantic interoperability, interoperability of the modelling phases is needed. Using the standardized levels of conceptual interoperability, modelling interoperability can be abstracted as *"pragmatic interoperability,"* i.e. methods' and procedures' awareness, lying above semantic interoperability [105]. Reps et al. recently introduced a standardized framework that leverages OHDSI and OMOP not only for *"transparently defining the problem and selecting suitable datasets,"* (i.e. semantics) but also for *"constructing variables from the observational data, learning the predictive model, and validating the model performance"* (i.e. modelling) [106].

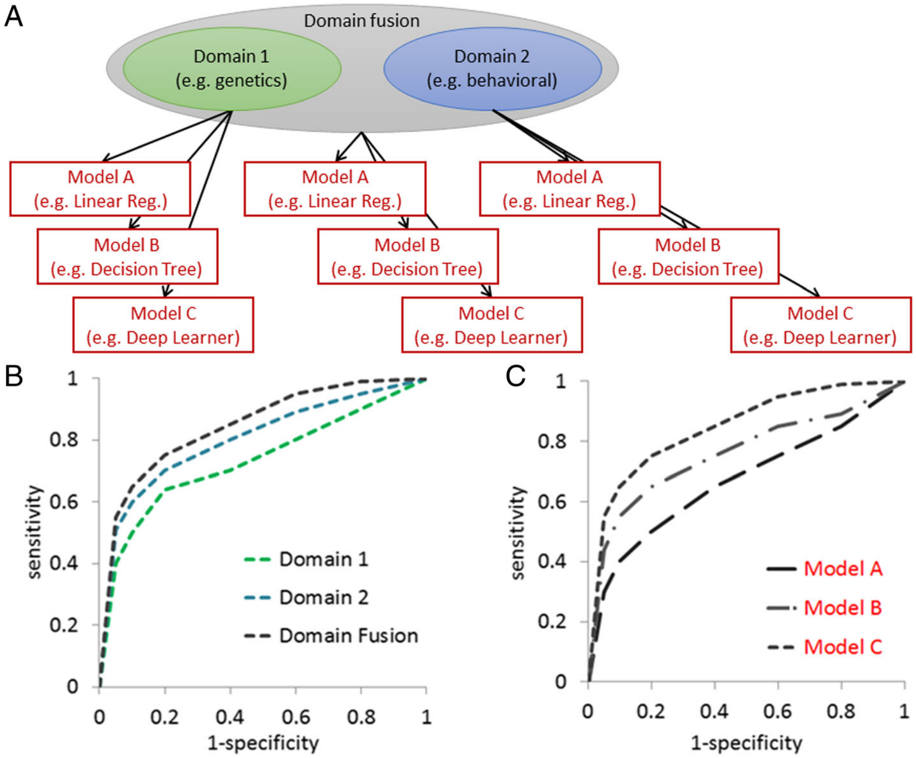


Fig. 7 Domain-guided and complexity-guided model selection. **a** Hypothetical data set with two domains and their merged domain, on which models of increasing complexity (linear regression, decision tree, and deep learner) are fit and compared. For example, using the receiver-operating characteristic, (**b**) the predictive performance of a prime model can be assessed using a single domain or merged domains, or (**c**) different models can be compared within the same domain space

Translational relevance

For any precision public health model to be useful, it should be robust to noise and generalizable; they should also be transparently presented in terms of their performance and reproducibility [107], and software libraries for differential privacy should be enforced as generic templates to facilitate data sharing and reproducibility of the works. When utilizing these models, we must consider whether the findings go beyond statistical significance and signify realms of clinical relevance. As previously mentioned, identification of risk factors which are immutable are impractical for interventions, and in cases of diseases where there are no treatments, accuracy of disease diagnoses will not impact clinical treatment decisions; however, additional insight on the mechanics of disease progression may be gained.

Linkage and systematization of data across multiple domains of life has the potential to increase patient education and participation in health care [108]. This in turn could lead to improvement in patient empowerment and shared decision-making, which are associated with improved health outcomes. By creating an access point for individuals to view their EHR and other variables that may affect their health, the health avatar empowers patients to take action. Impact of such empowerment has shown to modify health behaviors to reduce the risk of rheumatoid arthritis [109] and to make preparations for ill health in the future [110]. Moreover, health avatars can be venues for increased visibility of available health care facilities and ease of connection to care; this is currently being tested with wearable technology that can detect atrial fibrillation and prompt connection to a physician through mobile devices [111]. In addition to the impact on physicians for clinical decision support and on patient empowerment, the health avatar can be the missing intelligent algorithm that matches people to research studies based on their full medical history and other health-related factors. This will allow researchers to reach populations in vast numbers and allow implementation of novel study designs, such as examining rare adverse effects of a drug which randomized clinical trials cannot be sufficiently powered to detect [112].

The landscape of public health is evolving to a multi-domain, multi-stakeholder undertaking. The Food and Drug Administration is piloting digital health software programs. Companies which are outside of the health care domain are now engaged in creating health care programs for their employees [113].

However, a number of basic hurdles still remain open: prediction models of future health statuses are not yet accurate, and their actionability, i.e. changing the odds that a disease will occur, is even less accounted for; precision public health lack of contextualization within a societal and ecological environment; and integration with ethics and policymaking. Finally, affordability, trust, and education of the masses to this new paradigm of medicine will need to be addressed soon.

Conclusions

In this work, we have discussed the promises of precision medicine and precision public health, as well as the challenges we face to leverage big data for precision care that could lead to effective advancements and translational implementations. Thus, the aim of this paper was to provide a critical and objective view of where we are, and the work that needs to be done to achieve true precision medicine and precision public health, to improve health outcomes, and to reduce health disparities. In particular, we have revisited some of the definitions and described a hybrid theory-based and data-driven approach that can aid with the processes of study design and model inference. A hybrid approach allows us to tailor the modelling to specific problems and needs. The top-down approach relies on strong prior knowledge, which can be used to guide study design (e.g. domain selection, observational units, cohort identification) and test specific hypotheses (such as in clinical trials). On the other hand, the bottom-up approach helps in exploring a large variety of hypotheses with weaker assumptions.

Precision medicine demands interdisciplinary expertise that understands and bridges multiple disciplines and domains up to a point where the fulcrum of the research is located on the bridges themselves. This defines *transdisciplinarity*, knowledge discovery going beyond disciplines, which demands new research and development paradigms.

Acknowledgements
None.

Funding
This work was supported in part by: the "Virogenesis" project receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 634650; NIH grant UL1TR001427; the OneFlorida Cancer Control Alliance (funded by James and Esther King Biomedical Research Program, Florida Department of Health Grant Number 4 KB16); the OneFlorida Clinical Research Consortium (CDRN-1501- 26692) funded by the Patient Centered Outcomes Research Institute (PCORI). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or PCORI.

Availability of data and materials
Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Authors' contributions
MP conceived the work. JB, JM, FM, and MP together wrote and reviewed the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests

MP and JB are members of the editorial board of BMC Medical Informatics and Decision Making. The authors declare that they have no other competing interests.

Publisher's Note

Author details

[1]Department of Epidemiology, College of Medicine & College of Public Health and Health Professions, University of Florida, Gainesville, FL 32610, USA. [2]Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL 32610, USA. [3]Center for Health Outcomes and Informatics Research, Loyola University Chicago, Maywood, IL 60153, USA.

Received: 15 May 2018 Accepted: 4 December 2018
Published online: 29 December 2018

References

1. The Precision Medicine Initiative https://obamawhitehouse.archives.gov/precision-medicine. Accessed 12 Dec 2018.
2. Kohane IS. HEALTH CARE POLICY. Ten things we have to do to achieve precision medicine. Science. 2015;349(6243):37–8.
3. Adams SA, Petersen C. Precision medicine: opportunities, possibilities, and challenges for patients and providers. J Am Med Inform Assoc. 2016;23(4):787–90.
4. The Shift From Personalized Medicine to Precision Medicine and Precision Public Health: Words Matter! [https://blogs.cdc.gov/genomics/2016/04/21/shift]. Accessed 12 Dec 2018.
5. Jameson JL, Longo DL. Precision medicine--personalized, problematic, and promising. N Engl J Med. 2015;372(23):2229–34.
6. Barker RW. Is precision medicine the future of healthcare? Per Med. 2017;14(6):459–61.
7. Schork NJ. Personalized medicine: time for one-person trials. Nature. 2015;520(7549):609–11.
8. Ned RM. Genetic testing for CYP450 polymorphisms to predict response to clopidogrel: current evidence and test availability. Application: pharmacogenomics. PLoS Curr. 2010;2. https://doi.org/10.1371/currents.RRN1180.
9. Cargnin S, Jommi C, Canonico PL, Genazzani AA, Terrazzino S. Diagnostic accuracy of HLA-B*57:01 screening for the prediction of abacavir hypersensitivity and clinical utility of the test: a meta-analytic review. Pharmacogenomics. 2014;15(7):963–76.
10. Smith GD, Ebrahim S. Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol. 2003;32(1):1–22.
11. Hussain-Gambles M, Atkin K, Leese B. Why ethnic minority groups are under-represented in clinical trials: a review of the literature. Health Soc Care Community. 2004;12(5):382–8.
12. Liu KA, Mager NA. Women's involvement in clinical trials: historical perspective and future implications. Pharm Pract. 2016;14(1):708.
13. Vassiliou A, Vlastarakos PV, Maragoudakis P, Candiloros D, Nikolopoulos TP. Meniere's disease: still a mystery disease with difficult differential diagnosis. Ann Indian Acad Neurol. 2011;14(1):12–8.
14. Robinson PN. Deep phenotyping for precision medicine. Hum Mutat. 2012;33(5):777–80.
15. Delude CM. Deep phenotyping: the details of disease. Nature. 2015;527(7576):S14–5.
16. Dean K, Niven R. Asthma phenotypes and Endotypes: implications for personalised therapy. BioDrugs. 2017;31(5):393–408.
17. Castrillo JI, Oliver SG. Alzheimer's as a systems-level disease involving the interplay of multiple cellular networks. Methods Mol Biol. 2016;1303:3–48.
18. Naveed M, Ayday E, Clayton EW, Fellay J, Gunter CA, Hubaux JP, Malin BA, Wang X. Privacy in the genomic era. ACM Comput Surv. 2015;48(1):6.
19. Rehm HL. Evolving health care through personal genomics. Nat Rev Genet. 2017;18(4):259–67.
20. Krier J, Barfield R, Green RC, Kraft P. Reclassification of genetic-based risk predictions as GWAS data accumulate. Genome Med. 2016;8(1):20.
21. Paternoster L, Tilling K, Davey Smith G. Genetic epidemiology and Mendelian randomization for informing disease therapeutics: conceptual and methodological challenges. PLoS Genet. 2017;13(10):e1006944.
22. Arking D, Rommens J. Editorial overview: molecular and genetic bases of disease: enter the post-GWAS era. Curr Opin Genet Dev. 2015;33:77–9.
23. Kohane I. Deeper, longer phenotyping to accelerate the discovery of the genetic architectures of diseases. Genome Biol. 2014;15(5):115.
24. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747–53.
25. Dettling M, Gabrielson E, Giovanni P. Searching for differentially expressed gene combinations. Genome Biol. 2005;6(10):R88.
26. Graim K, Liu TT, Achrol AS, Paull EO, Newton Y, Chang SD, GR H, Cordero SP, Rubin DL, Stuart JM. Revealing cancer subtypes with higher-order correlations applied to imaging and omics data. BMC Med Genet. 2017;10(1):20.
27. de Vlaming R, Okbay A, Rietveld CA, Johannesson M, Magnusson PK, Uitterlinden AG, van Rooij FJ, Hofman A, Groenen PJ, Thurik AR, et al. Meta-GWAS accuracy and power (MetaGAP) calculator shows that hiding heritability is partially due to imperfect genetic correlations across studies. PLoS Genet. 2017;13(1):e1006495.
28. Ren Y, Gerke T, Kahveci T. Searching jointly correlated gene combinations. In: Unpublished work; 2017.
29. Huang B, Jiang C, Zhang R. Epigenetics: the language of the cell? Epigenomics. 2014;6(1):73–88.
30. Mensaert K, Denil S, Trooskens G, Van Criekinge W, Thas O, De Meyer T. Next-generation technologies and data analytical approaches for epigenomics. Environ Mol Mutagen. 2014;55(3):155–70.
31. Gligorijevic V, Malod-Dognin N, Przulj N. Integrative methods for analyzing big data in precision medicine. Proteomics. 2016;16(5):741–58.
32. Flanagan JM. Epigenome-wide association studies (EWAS): past, present, and future. Methods Mol Biol. 2015;1238:51–63.
33. Gilbert JA, Quinn RA, Debelius J, Xu ZZ, Morton J, Garg N, Jansson JK, Dorrestein PC, Knight R. Microbiome-wide association studies link dynamic microbial consortia to disease. Nature. 2016;535(7610):94–103.
34. Patel CJ, Bhattacharya J, Butte AJ. An environment-wide association study (EWAS) on type 2 diabetes mellitus. PLoS One. 2010;5(5):e10746.
35. Denny JC, Bastarache L, Roden DM. Phenome-wide association studies as a tool to advance precision medicine. Annu Rev Genomics Hum Genet. 2016;17:353–73.
36. Cusanovich DA, Caliskan M, Billstrand C, Michelini K, Chavarria C, De Leon S, Mitrano A, Lewellyn N, Elias JA, Chupp GL, et al. Integrated analyses of gene expression and genetic association studies in a founder population. Hum Mol Genet. 2016;25(10):2104–12.
37. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017;18(1):83.
38. Huang S, Chaudhary K, Garmire LX. More is better: recent Progress in multi-omics data integration methods. Front Genet. 2017;8:84.
39. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, Milanesi L. Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinformatics. 2016;17(Suppl 2):15.
40. Schroeder SA, Lecture S. We can do better--improving the health of the American people. N Engl J Med. 2007;357(12):1221–8.
41. Reading MJ, Merrill JA. Converging and diverging needs between patients and providers who are collecting and using patient-generated health data: an integrative review. J Am Med Inform Assoc. 2018;25(6):759–71.
42. Jain SH, Powers BW, Hawkins JB, Brownstein JS. The digital phenotype. Nat Biotechnol. 2015;33(5):462–3.
43. Holmberg C, J EC, Hillman T, Berg C. Adolescents' presentation of food in social media: an explorative study. Appetite. 2016;99:121–9.
44. Reece AG, Danforth CM. Instagram photos reveal predictive markers of depression. Epj Data Sci. 2017;6:15.
45. Dwyer-Lindgren L, Bertozzi-Villa A, Stubbs RW, et al. Inequalities in life expectancy among us counties, 1980 to 2014: temporal trends and key drivers. JAMA Intern Med. 2017;177(7):1003–11.
46. Lazer D, Kennedy R, King G, Vespignani A. The parable of Google flu: traps in big data analysis. Science. 2014;343(6176):1203–5.
47. Butler D. When Google got flu wrong. Nature. 2013;494(7436):155–6.
48. Buchan I, Winn J, Bishop C. A unified modeling approach to data-intensive healthcare. In: The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond: Microsoft Research; 2009. p. 91–8.
49. Maniadi E, Kondylakis H, Spanakis EG, Spanakis M, Tsiknakis M, Marias K, Dong F. Designing a digital patient avatar in the context of the

MyHealthAvatar project initiative. In: 13th IEEE international conference on BioInformatics and BioEngineering: 10–13 Nov. 2013 2013; 2013. p. 1–4.

50. Kim JH. Health avatar: an informatics platform for personal and private big data. Healthc Inform Res. 2014;20(1):1–2.
51. Epic. [http://www.epic.com]. Accessed 12 Dec 2018.
52. Cerner. [https://www.cerner.com]. Accessed 12 Dec 2018.
53. Belhajjame K, Zhao J, Garijo D, Gamble M, Hettne K, Palma R, Mina E, Corcho O, Gómez-Pérez JM, Bechhofer S, et al. Using a suite of ontologies for preserving workflow-centric research objects. Web Semant. 2015; 32(Supplement C):16–42.
54. Suarez A, Lutsko JF. Globally optimal fuzzy decision trees for classification and regression. Ieee T Pattern Anal. 1999;21(12):1297–311.
55. Harris PA, Scott KW, Lebo L, Hassan N, Lightner C, Pulley J. ResearchMatch: a national registry to recruit volunteers for clinical research. Acad Med. 2012; 87(1):66–73.
56. Feero WG, Wicklund CA, Veenstra D. Precision medicine, genome sequencing, and improved population health. JAMA. 2018;319(19):1979–80.
57. Sieverink F, Siemons L, Braakman-Jansen A, van Gemert-Pijnen L. Internet of things & personalized healthcare. Stud Health Technol Inform. 2016;221:129.
58. Showell C. Risk and the internet of things: Damocles, Pythia, or Pandora? Stud Health Technol Inform. 2016;221:31–5.
59. Lin C, Song ZH, Song HB, Zhou YH, Wang Y, Wu GW. Differential privacy preserving in big data analytics for connected health. J Med Syst. 2016;40:97.
60. Gruska DP. Differential privacy and security. Fund Inform. 2015;143(1–2):73–87.
61. Ebadi H, Sands D, Schneider G. Differential privacy: now it's getting personal. ACM SIGPLAN Not. 2015;50(1):69–81.
62. Dwork C, Roth A. The algorithmic foundations of differential privacy. Found Trends Theor Comput Sci. 2014;9(3–4):211–407.
63. Precision Public Health and Precision Medicine: Two Peas in a Pod [https://blogs.cdc.gov/genomics/2015/03/02/precision-public]. Accessed 12 Dec 2018.
64. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. J Med Internet Res. 2009;11(1):e11.
65. Vaithinathan AG, Asokan V. Public health and precision medicine share a goal. J Evid Based Med. 2017;10(2):76–80.
66. Khoury MJ, Bowen MS, Clyne M, Dotson WD, Gwinn ML, Green RF, Kolor K, Rodriguez JL, Wulf A, Yu W. From public health genomics to precision public health: a 20-year journey. Genet Med. 2017;20(6):574–82.
67. Golding N, Burstein R, Longbottom J, Browne AJ, Fullman N, Osgood-Zimmerman A, Earl L, Bhatt S, Cameron E, Casey DC, et al. Mapping under-5 and neonatal mortality in Africa, 2000-15: a baseline analysis for the sustainable development goals. Lancet. 2017;390(10108):2171–82.
68. The 42 V's of Big Data and Data Science [https://www.elderresearch.com/company/blog/42-v-of-big-data]. Accessed 12 Dec 2018.
69. Laney D. 3D data management: controlling data volume, velocity, and variety. In: META Group; 2001.
70. Khoury MJ, Galea S. Will precision medicine improve population health? JAMA. 2016;316(13):1357–8.
71. Gottlieb LM, Francis DE, Beck AF. Uses and misuses of patient- and neighborhood-level social determinants of health data. Perm J. 2018;22:18–078.
72. Matney SA. Semantic interoperability: the good, the bad, and the ugly. Nursing. 2016;46(10):23–4.
73. Marco-Ruiz L, Bellika JG. Semantic interoperability in clinical decision support systems: a systematic review. Stud Health Technol Inform. 2015; 216:958.
74. Liyanage H, Krause P, De Lusignan S. Using ontologies to improve semantic interoperability in health data. J Innov Health Inform. 2015; 22(2):309–15.
75. Bhatt M, Rahayu W, Soni SP, Wouters C. Ontology driven semantic profiling and retrieval in medical information systems. J Web Semant. 2009;7(4):317–31.
76. Codina L, Pedraza-Jimenez R. Ontologies and thesauri in information systems. Prof Inform. 2011;20(5):555–63.
77. World Health Organization's International Classification of Diseases (ICD) [http://www.who.int/classifications/icd/factsheet/en]. Accessed 12 Dec 2018.
78. SNOMED Clinical Terms (CT) [https://www.snomed.org/snomed-ct]. Accessed 12 Dec 2018.
79. Noy NF. Semantic integration: a survey of ontology-based approaches. Sigmod Rec. 2004;33(4):65–70.
80. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc. 2012;19(1):54–60.
81. He Z, Geller J, Chen Y. A comparative analysis of the density of the SNOMED CT conceptual content for semantic harmonization. Artif Intell Med. 2015;64(1):29–40.
82. Moreno-Conde A, Moner D, Cruz WD, Santos MR, Maldonado JA, Robles M, Kalra D. Clinical information modeling processes for semantic interoperability of electronic health records: systematic review and inductive analysis. J Am Med Inform Assoc. 2015;22(4):925–34.
83. Arvanitis TN. Semantic interoperability in healthcare. Stud Health Technol Inform. 2014;202:5–8.
84. Dentler K, ten Teije A, Cornet R, de Keizer N. Semantic integration of patient data and quality indicators based on openEHR archetypes. In: Process support and knowledge representation in health care: 2013//, vol. 2013. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013. p. 85–97.
85. Custovic A, Ainsworth J, Arshad H, Bishop C, Buchan I, Cullinan P, Devereux G, Henderson J, Holloway J, Roberts G, et al. The study team for early life asthma research (STELAR) consortium 'Asthma e-lab': team science bringing data, methods and investigators together. Thorax. 2015;70(8):799–801.
86. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, Suchard MA, Park RW, Wong IC, Rijnbeek PR, et al. Observational health data sciences and informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform. 2015;216:574–8.
87. Landers RN, Brusso RC, Cavanaugh KJ, Collmus AB. A primer on theory-driven web scraping: automatic extraction of big data from the internet for use in psychological research. Psychol Methods. 2016;21(4):475–92.
88. Lade SJ, Niiranen S. Generalized modeling of empirical social-ecological systems. Nat Resour Model. 2017;30(3):e12129.
89. Bizouarn P. Kenneth J. Rothman and multicausality in epidemiology. Rev Epidemiol Sante. 2012;60(1):59–69.
90. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, Liu PJ, Liu X, Marcus J, Sun M, et al. Scalable and accurate deep learning with electronic health records. npj Digit Med. 2018;1(1):18.
91. Shickel B, Tighe PJ, Bihorac A, Rashidi P, Deep EHR. A survey of recent advances in Deep learning techniques for electronic health record (EHR) analysis. IEEE J Biomed Health Inform. 2017;22(5):1589–604.
92. Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J, Doctor AI. Predicting clinical events via recurrent neural networks. JMLR Workshop Conf Proc. 2016;56:301–18.
93. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci Rep. 2016;6:26094.
94. Knight W. Biased algorithms are every where, and no one seems to care. In: MIT Technology Review. [https://www.technologyreview.com/s/608248/biased-algorithms-are-everywhere-and-no-one-seems-to-care/]. Accessed 12 Dec 2018.
95. Skeem JL, Lowenkamp CT. Risk, race, and recidivism: predictive Bias and disparate impact. Criminology. 2016;54(4):680–712.
96. Wu X, Zhang X. Automated inference on criminality using face images. https://arxiv.org/pdf/1611.04135v1.pdf. Accessed 12 Dec 2018.
97. Angwin J, Larson J, Mattu S, Kirchner L. Machine Bias; 2016.
98. Krause J, Perer A, Ng K. Interacting with predictions: visual inspection of black-box machine learning models. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. San Jose, California: ACM; 2016. p. 5686–97.
99. Fraccaro P, Nicolo M, Bonetto M, Giacomini M, Weller P, Traverso CE, Prosperi M, OS D. Combining macula clinical signs and patient characteristics for age-related macular degeneration diagnosis: a machine learning approach. BMC Ophthalmol. 2015;15:10.
100. van der Laan MJ, Polley EC, Hubbard AE. Super learner. Stat Appl Genet Mol Biol. 2007;6:Article25.
101. Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ. Mortality prediction in intensive care units with the super ICU learner algorithm (SICULA): a population-based study. Lancet Respir Med. 2015;3(1):42–52.
102. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics. 2007;8:25.
103. Olden JD, Jackson DA. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. Ecol Model. 2002;154(1–2):135–50.

104. Vapnik VN, Chervonenkis AY: On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. In: Measures of Complexity: Festschrift for Alexey Chervonenkis. Edited by Vovk V, Papadopoulos H, Gammerman A. Cham: Springer International Publishing; 2015: 11–30.
105. Robkin M, Weininger S, Preciado B, Goldman J. Levels of conceptual interoperability model for healthcare framework for safe medical device interoperability. In: 2015 IEEE symposium on product compliance engineering (ISPCE): 18–20 May 2015 2015; 2015. p. 1–8.
106. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc. 2018;25(8):969–75.
107. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). Ann Intern Med. 2015;162(10):735–6.
108. Bourzac K. Participation: power to the patients. Nature. 2016;537(7619):S66–8.
109. Sparks JA, Iversen MD, Yu Z, Triedman NA, Prado MG, Miller Kroouze R, Kalia SS, Atkinson ML, Mody EA, Helfgott SM, et al. Disclosure of personalized rheumatoid arthritis risk using genetics, biomarkers, and lifestyle factors to motivate health behavior improvements: a randomized controlled trial. Arthritis Care Res. 2017;70(6):823–33.
110. Zick CD, Mathews CJ, Roberts JS, Cook-Deegan R, Pokorski RJ, Green RC. Genetic testing for Alzheimer's disease and its impact on insurance purchasing behavior. Health Aff. 2005;24(2):483–90.
111. The Apple Heart Study [http://med.stanford.edu/appleheartstudy.html]. Accessed 12 Dec 2018.
112. Yih WK, Lieu TA, Kulldorff M, Martin D, McMahill-Walraven CN, Platt R, Selvam N, Selvan M, Lee GM, Nguyen M. Intussusception risk after rotavirus vaccination in U.S. infants. N Engl J Med. 2014;370(6):503–12.
113. Amazon's cloud on the healthcare horizon. Nat Biotechnol. 2018;36(3):205.