BMC Medicine

DEBATE Open Access

From hype to reality: data science enabling personalized medicine

Holger Fröhlich[1,21*], Rudi Balling[2], Niko Beerenwinkel[3], Oliver Kohlbacher[4,22,23,24], Santosh Kumar[5], Thomas Lengauer[6], Marloes H. Maathuis[7], Yves Moreau[8], Susan A. Murphy[9], Teresa M. Przytycka[10], Michael Rebhan[11], Hannes Röst[12], Andreas Schuppert[13], Matthias Schwab[14,25], Rainer Spang[15], Daniel Stekhoven[16], Jimeng Sun[17], Andreas Weber[18], Daniel Ziemek[19] and Blaz Zupan[20]

Abstract

Background: Personalized, precision, P4, or stratified medicine is understood as a medical approach in which patients are stratified based on their disease subtype, risk, prognosis, or treatment response using specialized diagnostic tests. The key idea is to base medical decisions on individual patient characteristics, including molecular and behavioral biomarkers, rather than on population averages. Personalized medicine is deeply connected to and dependent on data science, specifically machine learning (often named Artificial Intelligence in the mainstream media). While during recent years there has been a lot of enthusiasm about the potential of 'big data' and machine learning-based solutions, there exist only few examples that impact current clinical practice. The lack of impact on clinical practice can largely be attributed to insufficient performance of predictive models, difficulties to interpret complex model predictions, and lack of validation via prospective clinical trials that demonstrate a clear benefit compared to the standard of care. In this paper, we review the potential of state-of-the-art data science approaches for personalized medicine, discuss open challenges, and highlight directions that may help to overcome them in the future.

Conclusions: There is a need for an interdisciplinary effort, including data scientists, physicians, patient advocates, regulatory agencies, and health insurance organizations. Partially unrealistic expectations and concerns about data science-based solutions need to be better managed. In parallel, computational methods must advance more to provide direct benefit to clinical practice.

Keywords: Personalized medicine, Precision medicine, Stratified medicine, P4 medicine, Machine learning, Artificial intelligence, Big data, Biomarkers

Background

Personalized, precision, P4, or stratified medicine is understood as a medical approach in which patients are stratified based on their disease subtype, risk, prognosis, or treatment response using specialized diagnostic tests [1]. In many publications, the terms mentioned above are used interchangeably, although some authors make further distinctions between them to highlight particular nuances. The key idea is to base medical decisions on individual patient characteristics (including biomarkers) rather than on averages over a whole population. In agreement with the US Food and Drug Administration (FDA; https://www.fda.gov/ucm/groups/fdagov-public/@fdagov-drugs-gen/documents/document/ucm533161.pdf), we herein use the term biomarker for any measurable quantity or score that can be used as a basis to stratify patients (e.g., genomic alterations, molecular markers, disease severity scores, lifestyle characteristics, etc). The advantages of personalized medicine (summarized in [2, 3]) are widely considered to be (1) better medication effectiveness, since treatments are tailored to patient characteristics, e.g., genetic profile; (2) reduction of adverse event risks through avoidance of therapies showing no clear positive effect on the disease, while at the same time exhibiting (partially unavoidable) negative side effects; (3) lower healthcare costs as a

* Correspondence: holger.froehlich@ucb.com
[1] UCB Biosciences GmbH, Alfred-Nobel-Str. Str. 10, 40789 Monheim, Germany
[21] University of Bonn, Bonn-Aachen International Center for IT, Endenicher Allee 19c, 53115 Bonn, Germany
Full list of author information is available at the end of the article

consequence of optimized and effective use of therapies; (4) early disease diagnosis and prevention by using molecular and non-molecular biomarkers; (5) improved disease management with the help of wearable sensors and mobile health applications; and (6) smarter design of clinical trials due to selection of likely responders at baseline.

At present, personalized medicine is only an emerging reality. Molecular tumor boards at hospitals are probably furthest in realizing the promises of personalized medicine in clinical practice (Fig. 1). At the same time, this example already demonstrates a strong dependency of personalized medicine on computational solutions. Herein, we first explain, how modern approaches from data science, and specifically machine learning, are now beginning to impact personalized medicine. However, the way in which machine learning (often used interchangeably with the term Artificial Intelligence) is presented in the mainstream media often constitutes a hype, which must be contrasted with reality. We identify several challenges that currently constitute hurdles for realizing machine learning-based solutions more broadly in clinical practice. We discuss these challenges together with the existing potential of data science for personalized medicine. Finally, we highlight directions for future development.

Data science increasingly impacts personalized medicine

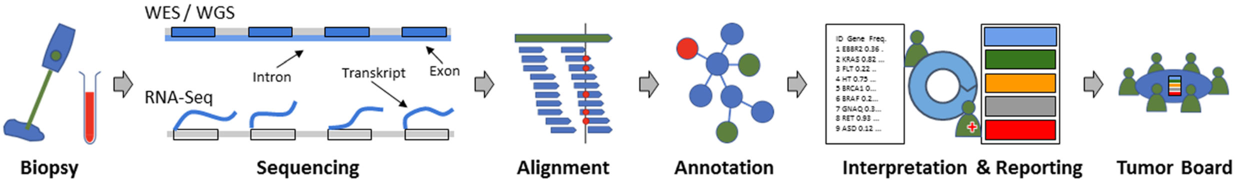
To date, the FDA has listed more than 160 (mostly genomic) pharmacogenomic biomarkers (https://www.fda.gov/Drugs/ScienceResearch/ucm572698.htm) and biomarker signatures (oncology: 33.5%; neurology: 6.1%) that have been approved for stratifying patients for drug response. For example, the anti-cancer drug trastuzumab (Herceptin®) can only be administered if the HER2/neu receptor is overexpressed because the drug interferes with this receptor. Personalized medicine is nowadays thus tightly connected with genomics. However, genomics and other biological high throughput data (transcriptomics, epigenomics, proteomics, metabolomics) are by no means the only source of data employed in the personalized medicine field. Other relevant data include, for example, bio-images (e.g., MRT and CT scans), electronic medical records (EMRs) [4], health claims data from insurance companies [5], and data from wearable sensors and mobile health applications [6].

It is important to mention that, in many cases, it is impossible to identify a single stratification factor or biomarker for patient populations. This is because many diseases (including cancer and various neurological and immunological diseases) are complex and affect a multitude of biological sub-systems. Accordingly, drugs for treating these diseases often target multiple proteins and

Box 1: Swiss Molecular Tumor Board

Molecular-based precision oncology is an emerging practice to improve cancer treatment by decreasing the risk of choosing treatments causing adverse events or lacking effectiveness (Shin et al., Precision Oncology 2017). The challenges of integrating molecular profiling with drug-gene interactions and clinical trial opportunities are manifold, but overcoming these is very promising (Hyman et al., Cell 2017). Furthermore, the clinical application does not only require short analysis turnaround time and the necessity of standardized high-quality workflows, but also requires delivery of the necessary evidence for clinicians to make an informed decision on the treatment of a patient.

The ETH Zurich technology platform NEXUS Personalized Health Technologies has established a workflow for cancer diagnostics based on comprehensive high-throughput sequencing of tumor samples (Singer et al., Bioinformatics 2017). The sequencing data are extensively analyzed using an algorithm, which combines several variant callers, queries multiple databases for annotation and puts individual variant frequencies and gene expression levels into the context of large cohorts. This allows for finding recruiting clinical trials, which potentially provide access to investigational treatments. The resulting treatment recommendations are summarized in a clinical report and discussed in the molecular tumor board at the clinic. This committee makes the final decision on the treatment of the patient.



The interpretation of the outputs of the pipeline are facilitated by an iterative process including clinicians to choose the most promising candidates from the automatically generated pipeline results. Selected candidate treatments are continuously recorded and included in future reporting. Moreover, NEXUS together with collaborators, is exploring text mining approaches to improve finding the most relevance literature evidence, given molecular and clinical information (Pasche et al., TREC Proceedings 2018).

Fig. 1 The Swiss molecular tumor board as an example of individualized, biomarker-based medical decisions in clinical practice

associated biological processes [7]. In general, clinical drug response is highly multi-faceted and dependent on a combination of patient intrinsic (e.g., genomic, age, sex, co-medications, liver function) and extrinsic (e.g., alcohol consumption, diet, sunlight exposure) factors [8]. In conclusion, single-analyte biomarker patient stratification, such as in the Herceptin® example, is only possible in special cases.

An alternative to single-analyte biomarkers are multi-analyte signatures derived from complex, high-throughput data, which allow patient characterization in a much more holistic manner than single biomarkers. Identifying marker signatures is difficult and requires state-of-the-art approaches offered by data science. Specifically, multivariate stratification algorithms using techniques from the area of Artificial Intelligence (including machine learning) play an increasingly important role (Fig. 2). A highly-cited example is MammaPrint™, a prognostic test for breast cancer based on a 70-gene signature [9], which was approved by the FDA in 2007. MammaPrint™ produces a score from the weighted average of 70 measured genes, which is predictive for the development of distant metastases. The clinical utility of the addition of the MammaPrint™ signature compared to standard clinicopathological criteria has been recently shown in selecting patients for adjuvant chemotherapy [10]. Other examples are *Geno2pheno* [11, 12], which is a computational tool used in clinical practice to estimate the resistance of HIV to an individual drug and to combinatorial therapies based on viral genotype (Fig. 3), and a gene signature (S3 score) for prediction of prognosis in patients with clear cell renal cell carcinoma [13].

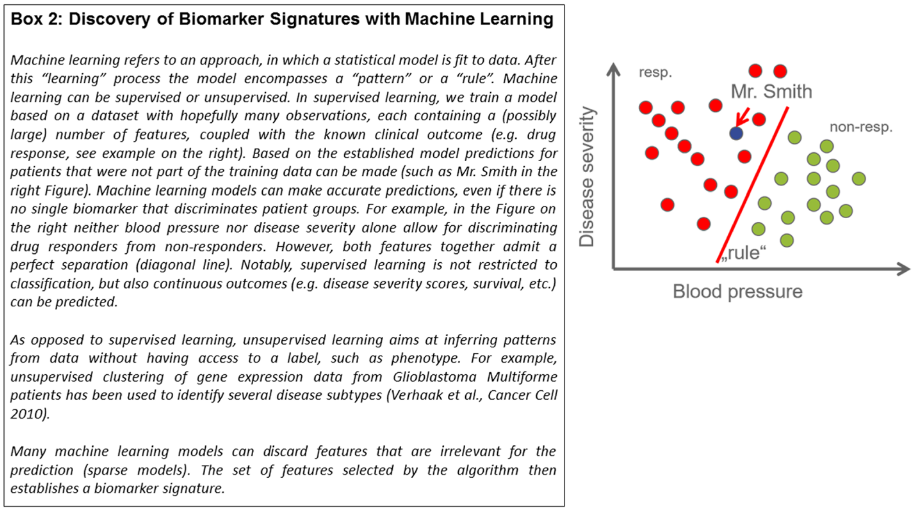
Driven by the increasing availability of large datasets, there is a growing interest into such data science-driven solutions. Specifically, 'deep learning' techniques have received a lot of attention, for example, in radiology [14, 15], histology [16] and, more recently, in the area of personalized medicine [17–20]. Some of these algorithms have been reported to achieve above-human diagnostic performance in certain cases [21]. Large commercial players now entering the field underline the widely perceived potential for machine learning-based solutions within personalized medicine (https://www.techemergence.com/machine-learning-in-pharma-medicine/, http://bigthink.com/ideafeed/for-new-era-of-personalized-medicine-google-to-store-individual-genomes-in-the-cloud, http://medicalfuturist.com/innovative-healthcare-companies/).

The data science and AI hype contrasts with reality

The mainstream media perception

From the previous discussion one might get the impression that enabling personalized medicine is mainly a matter of availability of 'big data', sufficient computing power, and modern deep-learning techniques. Indeed, this perception is portrayed in many mainstream publications, read by decision-makers in politics and industry (https://www.fool.com/investing/2017/09/21/3-ways-ai-is-changing-medicine.aspx, http://www.healthcareitnews.com/slideshow/how-ai-transforming-healthcare-and-solving-problems-2017?page=1, http://medicalfuturist.com/artificial-intelligence-will-redesign-healthcare/). In that context, some authors have even claimed the end of classical, hypothesis-driven science and stated that, in the future, all novel insights would come from an algorithmic analysis of large datasets (https://www.wired.com/2008/06/pb-theory/).

Such statements are overly optimistic and overlook several important aspects, which we discuss below.



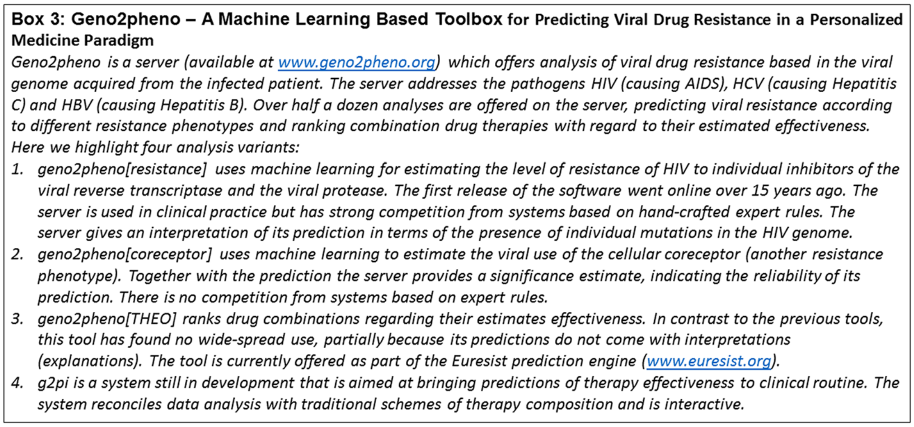
Box 2: Discovery of Biomarker Signatures with Machine Learning

Machine learning refers to an approach, in which a statistical model is fit to data. After this "learning" process the model encompasses a "pattern" or a "rule". Machine learning can be supervised or unsupervised. In supervised learning, we train a model based on a dataset with hopefully many observations, each containing a (possibly large) number of features, coupled with the known clinical outcome (e.g. drug response, see example on the right). Based on the established model predictions for patients that were not part of the training data can be made (such as Mr. Smith in the right Figure). Machine learning models can make accurate predictions, even if there is no single biomarker that discriminates patient groups. For example, in the Figure on the right neither blood pressure nor disease severity alone allow for discriminating drug responders from non-responders. However, both features together admit a perfect separation (diagonal line). Notably, supervised learning is not restricted to classification, but also continuous outcomes (e.g. disease severity scores, survival, etc.) can be predicted.

As opposed to supervised learning, unsupervised learning aims at inferring patterns from data without having access to a label, such as phenotype. For example, unsupervised clustering of gene expression data from Glioblastoma Multiforme patients has been used to identify several disease subtypes (Verhaak et al., Cancer Cell 2010).

Many machine learning models can discard features that are irrelevant for the prediction (sparse models). The set of features selected by the algorithm then establishes a biomarker signature.

Fig. 2 Discovery of biomarker signatures with machine learning



Box 3: Geno2pheno – A Machine Learning Based Toolbox for Predicting Viral Drug Resistance in a Personalized Medicine Paradigm

Geno2pheno is a server (available at www.geno2pheno.org) which offers analysis of viral drug resistance based in the viral genome acquired from the infected patient. The server addresses the pathogens HIV (causing AIDS), HCV (causing Hepatitis C) and HBV (causing Hepatitis B). Over half a dozen analyses are offered on the server, predicting viral resistance according to different resistance phenotypes and ranking combination drug therapies with regard to their estimated effectiveness.
Here we highlight four analysis variants:

1. *geno2pheno[resistance] uses machine learning for estimating the level of resistance of HIV to individual inhibitors of the viral reverse transcriptase and the viral protease. The first release of the software went online over 15 years ago. The server is used in clinical practice but has strong competition from systems based on hand-crafted expert rules. The server gives an interpretation of its prediction in terms of the presence of individual mutations in the HIV genome.*
2. *geno2pheno[coreceptor] uses machine learning to estimate the viral use of the cellular coreceptor (another resistance phenotype). Together with the prediction the server provides a significance estimate, indicating the reliability of its prediction. There is no competition from systems based on expert rules.*
3. *geno2pheno[THEO] ranks drug combinations regarding their estimates effectiveness. In contrast to the previous tools, this tool has found no wide-spread use, partially because its predictions do not come with interpretations (explanations). The tool is currently offered as part of the Euresist prediction engine (www.euresist.org).*
4. *g2pi is a system still in development that is aimed at bringing predictions of therapy effectiveness to clinical routine. The system reconciles data analysis with traditional schemes of therapy composition and is interactive.*

Fig. 3 Geno2pheno - a machine learning based toolbox for predicting viral drug resistance in a personalized medicine paradigm

Challenge 1: insufficient prediction performance for clinical practice

Machine learning methods capture and mathematically describe a (complex) signal that is present in a dataset. Their success does not only depend on the number of (patient) samples, but also on the signal-to-noise ratio. Indeed, separation of true signal from technical noise is still one of the key challenges in big data analysis [22] and one of the key aspects of any computational model. More generally, the prediction performance of any machine learning model is limited per se by the descriptive power of the employed data with respect to the clinical endpoint of interest. For example, EMRs are longitudinal, but largely phenotypic. Thus, molecular phenomena (e.g., non-common genomic variants) that might be relevant to stratifying patients are not sufficiently represented in the data. On the other hand, genomic data is mostly static (at least in non-cancerous tissues) and misses potentially important longitudinal clinical information. For each prediction problem, it is therefore critical to identify and combine the right data modalities that could contain parts of the relevant signal when starting to build machine learning models. Shortcomings can result in loss of prediction performance. Many machine learning models developed for personalized medicine do not have a predictive power close to the high (and potentially unrealistic) expectations of clinicians. Some of the reasons are as follows:

- The relationships of patient-specific characteristics to clinically relevant endpoints are highly complex and non-linear, often varying over time and, as mentioned before, typically not well described by one data instance alone. Furthermore, discriminating relevant from irrelevant patient-specific features remains a challenge, specifically in the field of biological high throughput (omics) data.
- It is challenging to obtain a sufficiently large patient cohort with well-defined phenotypes for training and testing models due to cost and time constraints.
- Many data (e.g., most omics data) are very noisy. There are two sources of this noise. One is technical measurement error (undesired), the other is biological variation (highly informative). We have no good methods for discriminating between these two kinds of noise.
- It can be challenging to quantitatively and objectively define clinical outcomes (e.g., in neurology, immunology, and psychology). This can lead to highly subjective and physician-dependent variations.
- Clinical outcomes may vary over time and be partially influenced by factors that are not patient intrinsic and thus hard to capture (e.g., social and environmental influences).
- A further factor impacting prediction performance is the careful choice of patient samples. Machine learning models are typically sensitive to selection biases, i.e., under- or over-represented specific patient subgroups in the training cohort, and there are currently under-explored ethical considerations at play as well. For example, over- or under-representation of certain ethnicities could result into a 'racist' prediction model [23]. A proper and careful design of the training set is necessary to ensure that it is representative for the population of patients in the intended application phase of the model in clinical practice.

Challenge 2: difficulties in interpretation

The scientific approach, which has been successfully established since the times of Galileo Galilei in the sixteenth century, always encompasses an ongoing process of hypothesis formulation and experimental validation [24]. While machine learning techniques can detect complex patterns in large data and provide accurate predictions, in general – we will discuss details later – they are unable to provide a deeper theoretical, mechanistic, or causal understanding of an observed phenomenon. Data science and AI thus do not replace classical, hypothesis-driven research. One reason is that machine learning models typically only capture statistical dependencies, such as correlation, from data. However, correlation does not imply causation. This is reflected by the fact that a multitude of biomarker signatures yielding similar prediction performance can be constructed to separate the same patient groups [25]. Even if an acceptable prediction performance can be achieved, the lack of a clear causal or mechanistic interpretation of machine learning models can hinder acceptance of data science-based solutions by physicians.

Challenge 3: insufficient validation for clinical practice

It is important to emphasize that establishing any algorithm for patient stratification in clinical practice requires rigorous validation. The quality of the fit of a sufficiently complex machine learning model to the training data (i.e., the training error) is usually highly over-optimistic and not indicative of its later performance on unseen data. A proper validation for clinical practice thus comprises several steps [10], as follows:

1. Internal validation based on the initial discovery cohort. This can be achieved by setting parts of the data aside as an independent test set or, more frequently, via cross-validation. Cross-validation refers to a strategy in which subsequently a certain fraction (e.g., 10%) of the original data is left out for model testing and the remaining part is used for model training. The cross-validation procedure averages prediction performance over different test sets and thus reduces the variance in test set performance estimates. This is specifically relevant if the overall discovery cohort is not very large.
2. External validation based on an independent cohort. This is necessary to address the potential selection bias during the compilation of the discovery cohort.
3. Validation in a prospective clinical trial to demonstrate the benefit compared to standard of care.

The entire process is time-consuming and costly. Consequently, the number of clinically validated models is limited.

Overall, the current hype about machine learning and AI in healthcare has to be contrasted with a number of existing challenges, which can be summarized as:

- Insufficient prediction performance
- Challenges with model interpretation
- Challenges with validation and translation of stratification algorithms into clinical practice

These challenges lead to the fact that, in contrast to the very high expectations portrayed in the mainstream media, there exist only very few examples of machine learning-based solutions that impact clinical practice (see the examples mentioned above). In the following, we discuss some of these challenges in more detail and point to possible ways of addressing them today and in the future.

What is possible today?

Machine learning for personalized medicine

Defining better clinical endpoints

Many methodological as well as applied articles focus on simple yes/no decision tasks, e.g., disease progression / no disease progression or clinical trial endpoint met /not met. This is surprising, because machine learning research offers a comprehensive arsenal of techniques to address clinical endpoints beyond binary classification, such as, real valued, time-to-event, multi-class or multivariate outcomes. Models with binary outcomes can be appropriate in specific situations, but in many cases, an appropriate clinical outcome is more complex. For instance, the commonly used response criterion for rheumatoid arthritis, a debilitating autoimmune disease of the joints, is based on the DAS28 disease score [26], which ranges on a continuous scale from 0 to 10 and is often discretized into three consecutive levels (low, medium, high disease activity).

The DAS28 score itself combines four components in a nonlinear equation, namely the number of swollen joints, the number of tender joints, plasma levels of CRP protein, and an assessment of the patient's global health as estimated by a physician. These components vary from discrete to continuous and from subjective, physician-dependent assessments to more objective measurements of biomarkers.

Another example is the prediction of response to anti-epileptic drug treatment. While at first glance overall seizure frequency reduction after a given number of weeks relative to baseline seems to be an appropriate endpoint in agreement to common practice in clinical trials, this choice in fact neglects the existence of different seizure types as well as the potential temporal modifications of these seizure types due to treatment. Thus, other and more complex (possibly multivariate) clinical endpoints might be necessary. We expect that a more careful choice of clinical endpoints as well as better technical monitoring capabilities (e.g., via mobile health applications and wearable sensors) will lead to more clinically useful prediction models in the future.

Defining appropriate model quality and performance measures

What makes a good model in personalized medicine? First, predictions must be accurate. As pointed out above, prediction accuracy must be assessed via a careful validation approach. Within such a validation procedure, it has to be decided how prediction performance will be measured. It appears that, in many studies, too much focus is given to standard, off-the-shelf metrics (e.g., area under the receiver operator characteristic curve), as compared to application-specific performance metrics. For instance, consider the case of predicting response to a first line therapy and assume that we can formulate this question as a classification task (responder vs. non-responder). Clearly, a perfectly accurate classifier is optimal. However, even a classifier that is mediocre with respect to overall accuracy might reliably identify those patients that will definitely not respond to the drug. The identified patients could immediately move on to a second line therapeutic and, thus, patient quality of life would improve and healthcare costs could be reduced. This example demonstrates the relevance of carefully defining appropriate prediction performance metrics.

However, prediction performance is only one aspect of judging the overall quality of a model. Another aspect is model stability, which reflects the degree to which a model (including variables selected by that model) remains the same if the training data is slightly changed. Model stability is a particular issue when working with gene expression data, where models trained on very different or even disjoint gene subsets can result in similar prediction performance regarding a given clinical endpoint, since highly correlated features can be substituted for each other [26]. Model stability should be routinely reported in addition to prediction performance.

Various methods have been developed for increasing the chance of obtaining a stable model during the development phase of a stratification algorithm. For example, inclusion of prior knowledge, such as biological networks and pathways, can enhance the stability and thus reproducibility of gene expression signatures [27–29]. Moreover, zero-sum regression [30] can be used to build classifiers that are less dependent on the employed omics platform (e.g., a specific microarray chip) [31], thus easing external validation, translation into clinical practice as well as long-term applicability of the model. We think that more frequent use of such methodology in conjunction with careful evaluation of model stability would lower the barrier for model transfer from discovery to external validation and finally to clinical application.

Tools for interpreting a machine learning model

As researchers collect and analyze increasingly larger sets of data, a greater number of sophisticated algorithms are employed to train predictive models. Some of the computational methods, in particular those based on deep learning techniques, are often criticized for being black boxes. Indeed, as the number of input features becomes large and the computational process more complex, understanding the reasons for obtaining a specific result is difficult, if not impossible. In many instances, for example, in the case of identification of disease markers, understanding the computational decision-making process leading to the selection of specific markers is, however, necessary and demanded by physicians. Using black-box models for medical decision-making is thus often considered to be problematic, leading to initiatives like the 'right to an explanation' law Article 22 of the General Data Protection Regulation propositioned by the European Union in April 2016/679. Similarly, in the process of drug development in pharmaceutical industry, regulatory agencies require transparency and supporting evidence of a molecular mechanism for the choice of specific biomarker panels.

While usefulness of data-driven prediction is increasingly recognized, a key requirement for credibility of such solutions is thus the ability to interpret them in the context of current biomedical knowledge. It is important to understand that the concept of interpretability covers a spectrum (Fig. 4). At one end of the spectrum, there is a detailed understanding of the exact (biochemical) molecular and pathophysiological mechanisms that link a model with a defined clinical endpoint. Typically, this level of insight is rarely achievable due to lack of knowledge.

A less detailed level of understanding is that of total causal effects of a predictor regarding the clinical endpoint of interest. For example, in a randomized controlled clinical trial, any difference in outcomes between the two treatment groups is known to be caused by the treatment (since the groups are similar in all other respects due to the randomization). Thus, although one may not know exactly how the treatment affects the outcome, one knows that it does. Such statements about total causal effects are more difficult to obtain in a setting outside clinical trials, where purely observational data from untreated patients are collected (e.g., cross-sectional gene expression data). Nonetheless, computational approaches have significantly advanced in this field over recent years and, under certain assumptions and conditions, allow for estimating causal effects directly from observational data [32, 33].

At a lower level of interpretability, gene set and molecular network analysis methods [34, 35] can help to understand the biological sub-systems in which biomarkers selected by a machine learning algorithm are involved. There also exists a large body of literature on

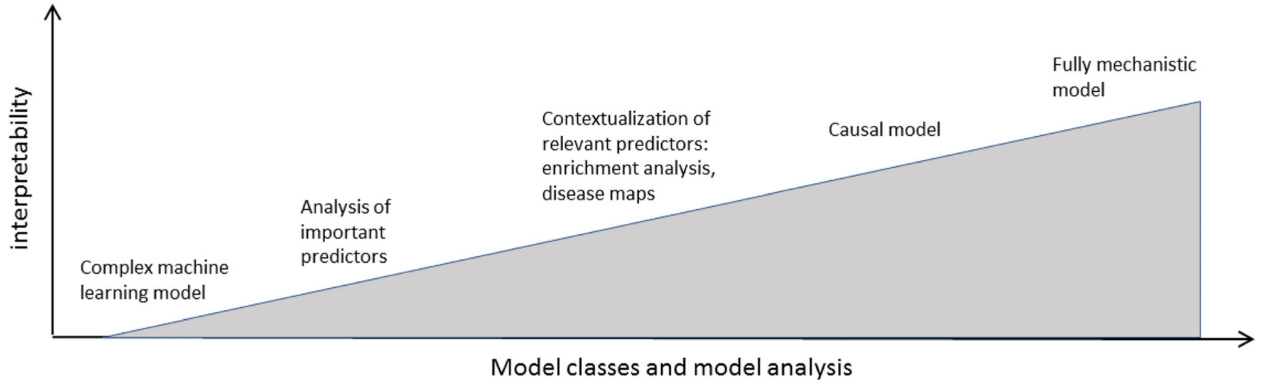

interpretability
Complex machine learning model
Analysis of important predictors
Contextualization of relevant predictors: enrichment analysis, disease maps
Causal model
Fully mechanistic model
Model classes and model analysis

Fig. 4 Different classes of machine learning models and their interpretability via model analysis

how to directly incorporate biological network information together with gene expression data into machine learning algorithms (see [28] for a review).

Recently, the concept of 'disease maps' has been developed as a community tool for bridging the gap between experimental biological and computational research [36]. A disease map is a visual, computer-tractable and standardized representation of literature-derived, disease-specific cause–effect relationships between genetic variants, genes, biological processes, clinical outcomes, or other entities of interest. Disease maps can be used to visualize prior knowledge and provide a platform that could help to understand predictors in a machine learning model in the context of disease pathogenesis, disease comorbidities and potential drug responses. A number of visual pathway editors, such as CellDesigner [37] and PathVisio [38], are used to display the content of a disease map and to offer tools for regular updating and deep annotation of knowledge repositories. In addition, dedicated tools such as MINERVA [39] and NaviCell [40] have been developed by the Disease Map community. At this point in time, disease maps are more knowledge management rather than simulation or modeling tools, although intensive efforts are underway to develop the next generation of disease maps that are useful for mathematical modelling and simulation and become an integral part of data interpretation pipelines.

The least detailed level of understanding of a complex machine learning algorithm is provided by the analysis of relative importance of variables with respect to model predictions. Relative variable importance can be calculated for a range of modern machine learning models (including deep learning techniques), but the level of insight depends on whether only few out of all variables have outstanding relevance and whether these variables can be contextualized with supporting evidence from the literature. It is also not clear a priori if such variables are only correlated with or perhaps also causal for the outcome of interest. Finally, inspecting most important variables may be less informative in the case of highly collinear dependencies among predictor variables such as, for example, in gene expression data.

In addition to the interpretation of predictors there is a need from a physician's perspective to better understand model predictions and outputs for a given patient. One obvious way might be to display patients with similar characteristics. However, the result will depend on the exact mathematical definition of similarity. Moreover, clinical outcomes of most similar patients will, in general, not always coincide with the predictions made by complex machine learning models, which could result in misinterpretations. The same general concern applies to approaches, in which a complex machine learning model is approximated by a simpler one to enhance interpretability, for example, using a decision tree [41, 42].

Data type-specific challenges and solutions

Real-world longitudinal data

Longitudinal EMR and claims data have received increasing interest in recent years within the field of personalized medicine [43, 44] since they provide a less biased view on patient trajectories than data from classical clinical trials, which are always subject to certain inclusion and exclusion criteria [45]. Specifically in the United States, a whole industry has grown to collect, annotate, and mine real-world longitudinal data (https://cancerlinq.org/about, https://truvenhealth.com/). The recent US$1.9 billion acquisition of Flatiron Health by the pharma company Roche (https://www.roche.com/media/store/releases/med-cor-2018-02-15.htm) marks the potential that is seen by industrial decision-makers in the context of drug development, pharmacovigilance, label expansion, and post-marketing analysis [45, 46].

Longitudinal real-world data pose specific challenges for training and validation of predictive models. Within the analysis of clinical real-world databases (e.g., Clinical

Practice Research Datalink; https://www.cprd.com/home/) patients for a study cohort are typically selected based on a specified index date or event, which is often difficult to define and thus leaves room for different choices. Since the maximal observation horizon in real-world databases is often limited to a certain number of years (e.g., due to budget restrictions), some patients are longer observed than others. Specifically, claims data may contain gaps (e.g., due to periods of unemployment of patients) and the exact date of a diagnosis, prescription, or medical procedure cannot be uniquely determined. It is not always clear for the treating physician which ICD diagnosis codes to choose, and this leaves room for optimization with respect to financial outcomes. In addition, EMRs require natural language preprocessing via text mining, which is a difficult and potentially error-prone procedure in itself. In conclusion, development of a predictive model for personalized medicine based on real-world clinical data thus remains a non-trivial challenge.

Classically, validation of a predictive model relies on an appropriate experimental design and randomization. Real-world data often limits the options available for rigorous validation. Classical strategies, such as carefully crafted cross-validation schemes, can offer reliable validation, but they might be tricky to design, and the limits of such retrospective validation must be properly understood. Another option is the use of different time windows where only retrospective data up to a given date is used to develop a model, which is then used on the data available after this date. Such a setup can be close to an actual prospective evaluation, although the risk for biases is larger. Another option is to consider such analyses as only generating hypotheses, which are then followed up in a more classical fashion by setting up a carefully designed observational study manifesting the final validation. A more speculative possibility is the adaptation of so-called A/B testing techniques that are common in web development and software engineering [47]. This would entail randomization of patients for therapeutic options directly in the real-world environment. While such a setting is probably not feasible for drug development, it may be applicable to determine the efficacy of interventions in a real-world setting or to determine the right patient population for a given intervention.

Multi-modal patient data

There is an increasing availability of multi-scale, multi-modal longitudinal patient data. Examples include the Alzheimer's Disease Neuroimaging Initiative (http://adni.loni.usc.edu/) (omics, neuro-imaging, longitudinal clinical data), the Parkinson's Progression Markers Initiative (http://www.ppmi-info.org/) (omics, neuro-imaging, longitudinal clinical data), the All-of-Us Cohort (https://allofus.nih.gov/) (omics, behavioral, EMRs, environmental data), the GENIE project (http://www.aacr.org/Research/Research/Pages/aacr-project-genie.aspx#.WvqxOPmLTmE) (genomic and longitudinal real-world clinical data) and, specifically for multi-omics, the NCI's Genomic Data Commons [48]. Multi-modal data provide unique opportunities for personalized medicine because they allow for capturing and understanding different dimensions of a patient. This aspect is in turn widely believed to be key for enhancing the prediction performance of stratification algorithms up to a level that is useful for clinical practice. Accordingly, there has been a lot of work in methods that combine data from different (omics-) modalities, see [49] for a review.

A major bottleneck in current studies collecting multiple data modalities of clinical cohorts is posed by the fact that different studies are often performed on cohorts of different patients and different experimental approaches are used across studies (see Fig. 5 for an example). As consequence, data from different studies becomes difficult or even impossible to integrate into a joint machine learning model. Several strategies are possible to reduce this problem in the future. A first strategy is to perform systematic multi-modal data assessment of each individual in a clinically rigorously characterized cohort, including longitudinal clinical and omics follow-up. In the more classical clinical setting, the success of the Framingham Heart Study (https://www.framinghamheartstudy.org/) comes to mind, which is a long-term study about risk factors for cardiovascular diseases running since 1948. While, in the future, we will analyze larger and larger volumes of real-world data, we should be aware of the limitations of such data (interoperability of data from different sources, non-systematically

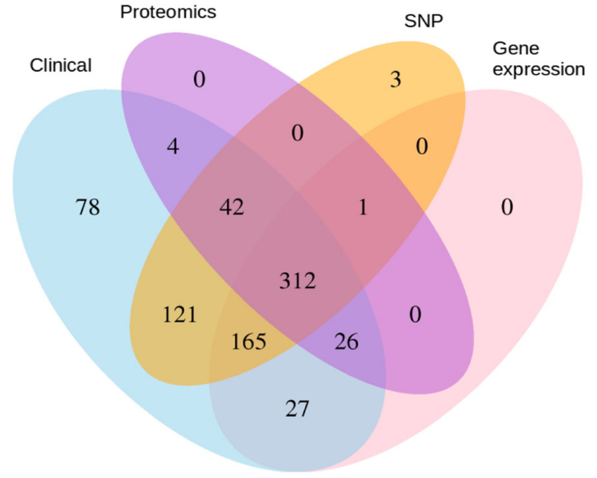


Fig. 5 Overlap of different omics data entities and clinical data in the AddNeuroMed Alzheimer's Disease cohort from EMIF-AD (http://www.emif.eu/about/emif-ad). Numbers refer to patients, for which a particular data modality is available

collected data, measurement quality, inconsistencies and errors, etc.). Rigorous multi-modal observational studies are essential for establishing reliable baselines for the development of real-world models. Ideally, multi-modal data would be collected longitudinally in regular intervals for all subjects. While this has been achieved for individual studies [50], for practical and economic reasons, this is likely to be limited to a small number of cohorts. A second approach is to have some overlap among patients across different cohorts. Statistical methods and machine learning can then be used to 'tie' different datasets together. A third approach is to collect a joint modality (such as standardized clinical data or biomarkers) across different studies. This joint modality again makes it possible to tie together different datasets. It must be stressed that this problem of disconnected cohorts is currently a major obstacle for leveraging multi-omics data.

It should be emphasized that, ideally, multi-modal, multi-omics data should be considered in conjunction with longitudinal clinical data. Despite of the examples mentioned above (Alzheimer's Disease Neuroimaging Initiative, Parkinson's Progression Markers Initiative, All-of-Us Cohort) we are currently just in the beginning of performing corresponding studies more systematically. The combination of multi-omics with real-world longitudinal data from clinical practice (e.g., EMRs) and mobile health applications marks a further potential for personalized medicine in the future. The GENIE project is an important step into this direction.

Translating stratification algorithms into clinical practice

The ability to accelerate innovation in patient treatment is linked to our ability to translate increasingly complex and multi-modal stratification algorithms from discovery to validation. Stratification in clinical application means assigning treatment specifications to a particular patient, which may include type, dosage, time point, access to the treatment, and other pharmacological aspects. The validation of such algorithms is usually performed via internal validation (cross-validation), external validation (using a separate patient cohort), and prospective clinical trials compared to the standard of care [10] (http://www.agendia.com/healthcare-professionals/the-mindact-trial/). Proper validation constitutes a requirement for translating these methods to settings in which they can generate impact on patient outcomes. In addition to classic healthcare providers, such as hospitals and general practitioner, mobile health applications and wearable sensors might play an increasing role in the future. As described earlier, integrating multi-modal data is key for gaining new insights and lies also at the heart of stratifying patients for diagnostic, predictive, or prognostic purposes. However, considerable barriers exist regarding the integration of similar data from different cohorts, normalization of data across measurement platforms, and the ability to process very large volumes of data in appropriate systems close to or within the clinical infrastructure remains limited. Strictly controlled cloud services, which appropriately protect patient data, could be an approach to alleviating this limitation [51]. At this point it might be possible to learn from organizations that today handle large scale real-world clinical data (mostly in the US). However, their approaches may have to be adapted to the legal environments in each specific country.

At present, translation of algorithms for patient stratification into clinical practice is also difficult due to regulatory aspects. Prospective clinical trials required for approval of diagnostic tools by regulatory agencies are very costly and the challenges for finding sponsors are high. One possibility of lowering the associated barriers might be to perform a stepwise approach with initial pilot studies to exemplify the value that can be gained for patients, healthcare sustainability, translational science, and economic efficiency. Such projects would need to showcase the principle value of patient stratification. Moreover, they could provide meaningful insights into disease biology (via biomarkers). These outcomes should ideally be measured longitudinally after machine learning-based stratification and thus provide a feedback loop that helps improve the stratification algorithm.

A commonly stated myth is that health innovation is based on the paradigm of build-and-freeze (https://www.theatlantic.com/technology/archive/2017/10/algorithms-future-of-health-care/543825/), which means that software is built, frozen, and then tested in unchanged form for its lifetime. However, development of better stratification algorithms will require a more seamless updating scheme. There have been interesting developments in recent years in terms of regulation and risk management for continuous learning systems. An example of such a development is the Digital Health Software Precertification (Pre-Cert) Program (https://www.fda.gov/MedicalDevices/DigitalHealth/DigitalHealthPreCertProgram/Default.htm) launched recently by the FDA. PreCert aims at learning and adapting its key elements based on the effectiveness of the program. In addition, Clinical Laboratory Improvement Amendments (CLIA; https://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/IVDRegulatoryAssistance/ucm124105.htm) labs provide a template for how health-related software tools developed to inform precision medicine can be validated in a clear and transparent manner as the tool is continually updated. CLIA labs are certified labs that go through a process of regular certifications monitored by the FDA and other regulatory agencies in the US. These labs are required to follow approved and documented Standard Operation Procedures. They can use medical devices, which can include software for diagnostics, given that they employ such Standard Operation Procedures and waive the certification process (https://wwwn.cdc.gov/clia/Resources/WaivedTests/default.aspx). Most importantly, the developer of the tool can update the software. The CLIA labs are independent in deciding whether they will re-validate the software and can adopt a strategy that best serves the technological pace of the software and their clinical needs with respect to increased capabilities or better performance. For instance, a lab may decide to validate only major version releases, such as going from version 1.x to 2.0, and have minor version releases included on the fly.

The vision of precision medicine is to provide the right intervention to the right patient, at the right time and dose. The described approaches, based on iterative feedback between the developers and the clinical end users, could increase our ability to adapt stratification algorithms better to new insights in disease biology, access to new molecular data, and changes in clinical settings. This has been a challenge with promising predictive models often failing validation in independent studies. Real-world longitudinal data from clinical practice and data collected through wearables or other means of participatory data collection cannot only widen the spectrum of possible data sources to build new stratification algorithms [52, 53], but they may also be partially included in clinical trials for validation purposes of stratification algorithms.

What could be possible tomorrow?

Novel approaches to better link prediction algorithms with biomedical knowledge

As discussed earlier, challenges with the interpretation of complex machine learning models are one of the important bottlenecks for applying personalized medicine more widely. Innovative software solutions are needed to better put complex machine learning models and their outputs into the context of computationally accessible knowledge of human pathophysiology.

While the current standard is to map the most relevant molecular features in a machine learning model onto biological pathways, this approach could be further enhanced to make machine learning-based decisions interpretable by clinicians. In the future, one might imagine software systems that automatically collect information on each variable from various databases and publications (e.g., via text mining). Such tools could eventually even compose entire reports (including supporting texts and figures of disease maps) for each individual feature in a machine learning model. Such reports could thus automatically contextualize each variable with the multitude of available biomedical knowledge in a fully interactive fashion. The physician could zoom and filter specific aspects of a model on demand.

Another idea is to visualize entire patient trajectories (originating, for example, from longitudinal clinical trials, real-world clinical or behavioral data) within interactive ‘disease landscapes’ (essentially low-dimensional data projections). Such a tool could help physicians to understand disease development over time. Taking the patient’s history into account will allow clinicians to visualize and interpret the speed and severity of disease progression. Individual patient trajectories could then be followed and compared to determine, for example, which intervention is appropriate for which patient and at what time [54]. Similar concepts have been developed in other contexts, e.g. for estimating the in-vivo fitness landscape experienced by HIV-1 under drug selective pressure [55].

The development of such methods and software systems will be a major effort and will likely require a substantial text analytical and software engineering component. However, such systems could greatly facilitate the communication between computational scientists and physicians and help make complex machine learning models more interpretable.

Going from ‘what’ to ‘why’ – towards better interpretable modeling approaches

Causal models

Machine learning models are typically neither mechanistic nor causal. They largely capture (non-linear) correlations between predictor variables and clinical outcomes and are thus often criticized for being black boxes. The main advantage of modern machine learning approaches is that they neither require a detailed prior understanding of cause–effect relationships nor of detailed mechanisms. The main limitation is the difficulty to interpret them (see previous Section). A major question thus relates to how far machine learning methods could evolve into more causal models in the future.

Causal graphical models (causal Bayesian networks in particular) constitute an established framework for causal reasoning [56]. They provide a compact mathematical and visual representation of a multivariate distribution, and more importantly, they allow to make predictions of the system under unseen interventions (e.g. a new treatment or a gene knockout). Under appropriate assumptions, causal graphical models can be learned from observational data [57–59]. In doing so, it is also possible to incorporate background knowledge or to allow for hidden or unmeasured confounders. We refer to [60] for a review paper.

Causal graph learning methods may play an increasingly important role in the future in identifying predictor variables with causal influence on clinical outcomes [61] and may thus help to move towards a causal interpretation of predictor variables in a machine learning model

[62]. However, there are non-trivial challenges that need to be addressed, such as dealing with violations of assumptions, high computational costs and non-linear relationships [63].

Hybrid machine learning and mechanistic models

Despite the increasing availability of massive datasets, the predictive power of most of the available disease models does not yet satisfy the requirements for clinical practice. One of the reasons is that, in principle, predictive disease models must cover all relevant biotic and abiotic mechanisms driving disease progression in individual patients. Although the primary disease-driving mechanisms are often aberrations at the molecular level, such as mutations in the genome, disease progression is affected by the robustness of the overall system. However, biological systems have established a multitude of repair mechanisms to compensate for the effects of molecular aberrations, thus introducing feedback loops and non-linear interactions into the system [64]. Overall, disease progression is a process affected by a multitude of highly diverse mechanisms across biological hierarchies, which are differently expressed in individual patients.

Thus, a disease model, designed for applications in precision medicine in clinics, must in principle integrate three conceptual layers:

- A core disease model (CDM) represents only the known intra- and inter-cellular processes that are the key drivers of the disease in an average patient.
- The CDM must be adapted to the individual patient and their specific medical history and environment, such as genetic variations, co-morbidities or physiology, by environment adaption models (EAM). The EAM must provide an individualization of the parameters controlling the CDM, eventually combined with an individualized re-structuring of the CDM, e.g., by adding or dropping biological mechanisms that are relevant only in specific patient populations.
- Monitoring models must be developed to describe how clinically accessible outcome measurements representing the disease evolution are linked to the CDM.

Today, fully mechanistic models exist for a series of disease-driving core processes at the molecular and cell population level [65]. However, broader application of mechanistic modelling to implement the CDM for complex diseases is hampered by insufficient knowledge of the interaction of the core disease-driving mechanisms across scales. Even worse, the relevant mechanisms for EAM and monitoring models are almost never completely known. Altogether, it thus seems unlikely that fully mechanistic models will play a dominant role in personalized medicine in the near future.

While machine learning models are not harmed by insufficient biomedical knowledge, they are often criticized for their black-box character. Hybrid modelling, also named grey-box or semi-parametric modelling, is an integrative approach combining available mechanistic and machine learning-based sub-models into a joint computational network. The nodes represent model components and the edges their interaction. First combinations of mechanistic and data-driven models have been developed for chemical and biotech process modelling [66, 67]. For example, neural networks have been used to compensate the systematic errors of insufficient mechanistic models, to estimate unobservable parameters in mechanistic models from observable data, or to estimate the interaction between different mechanistic sub-models [68, 69].

A further successful example of hybrid modeling comprises learning the drug mechanism of action from data [70, 71]. Hybrid models may thus be a way to combine the positive aspects of fully mechanistic and purely data-driven machine learning models. First showcases have demonstrated the potential, but more successful applications are needed. Moreover, a deeper understanding of the theoretical capabilities of hybrid models as well as their limitations is necessary.

Controlling critical transitions in patient trajectories

One of the key objectives of personalized medicine is predicting the risk of an individual person to develop a certain disease or, if the disease has already developed, to predict the most suitable therapy. This also includes predicting the likely course of disease progression. Disease trajectories entail all the hallmarks of a complex system. In this sense, modeling disease trajectories is not fundamentally different from attempts to model and simulate other complex systems such as the climatological, ecological, economic or social systems. In many of these highly nonlinear, complex systems with thousands or millions of components, involving redundant and intertwined feedback relations, so called critical transitions or catastrophic shifts can be observed. Such transitions are defined by critical thresholds, sometimes called tipping points at which a system transitions abruptly from one state to another, seem to exist. However, in many of these cases, critical transitions are extremely difficult to predict in advance.

For certain diseases, we believe that the concept of critical transitions might also be applicable in the context of personalized medicine. Tipping points are often observed during the course of acute or chronic disease development. The ability to predict a critical transition of a developing disease before it really happens would be highly desirable and provide very valuable pre-disease biomarkers.

Recently, Liu et al. [72] used gene expression analysis to develop the concept of dynamic network biomarkers, where higher-order statistical information is used to identify upcoming tipping points. The idea is that, during the disease trajectory, a subset of genes starts to fluctuate and leads to a destabilization of a (possibly high-dimensional) attractor state. By measuring the changes in gene correlation in addition to changes in the variation of gene expression, a quantitative index was proposed as an early warning signal for a critical transition.

Towards an evolutionary understanding of human disease

From a broader perspective, evolutionary principles could help to improve our understanding of human disease [73]. Evolutionarily conserved control genes are probably highly relevant for the proper functioning of molecular pathways [74], and evolutionary history of human disease genes reveals phenotypic connections and comorbidities among some diseases [75]. We are now at the verge of reconstructing the molecular and cellular circuitry of embryogenesis [76]. In addition, whole-genome next-generation sequencing efforts of hundreds of thousands and soon Millions of patients with common and rare diseases provide us with a rich genotype–phenotype landscape underlying the development and manifestation of human diseases. Such data provides interesting opportunities to better understand the influence of genomic variants on evolutionarily conserved genomic regions and molecular networks in the context of human diseases.

Evolutionary conservation might be relevant for constraining models and simulating human diseases. Biologically possible and plausible disease trajectories are likely limited by the topological and dynamic upper and lower bounds that are set by the evolutionary history of a disease network. A key challenge for personalized medicine is to come up with a mechanistic explanation of an individual's disease development. We need to understand the effects of genetic variation on the resulting phenotypic variation. This requires close cooperation between disciplines striving for an integration of the concepts of ontogeny and phylogeny. Human diseases must be seen in the light of evolution and models of human diseases need to integrate data, information, and knowledge from developmental biology and embryology.

Conclusions

In the era of growing data volumes and ever shrinking costs for data generation, storage, and computation, personalized medicine comes with high promises, which can only be realized with the help of advanced algorithms from data science, particularly machine learning. Modern machine learning algorithms have the potential of integrating multi-scale, multi-modal, and longitudinal patient data to make relatively accurate predictions, which, in some examples, may even exceed human performance [21]. Large commercial players that are now entering the field of medicine underline the potential that is widely seen for computational solutions.

However, the current hype around AI and machine learning must be contrasted with reality. While many prediction algorithms for patient stratification have been published over the last decade, only very few approaches have reached clinical practice so far. Major existing bottlenecks discussed in this paper include the (1) lack of sufficient prediction performance due to a lack of signals in the employed data; (2) challenges with model stability and interpretation; (3) a lack of validation of stratification algorithm via prospective clinical trials, which demonstrate benefit compared to standard of care; and (4) general difficulties to implement a continuous maintenance and updating scheme for decision support systems.

In addition, general concerns around data privacy as well as ethical and legal aspects must not be overlooked. To overcome these hurdles, an interdisciplinary effort including computational scientists, physicians, patient advocates, regulatory agencies, and health insurance providers is required in the context of a 'learning healthcare system' (http://www.learninghealthcareproject.org/section/background/learning-healthcare-system). There is a need to better manage the (partially unrealistic) expectations and concerns about data science and AI-based solutions.

In parallel, computational methods must advance in order to provide direct benefit to clinical practice. Current algorithms are far from being able to recommend the right treatment at the right time and dose for each patient. Steps that bring us closer to this goal could be (1) innovative software tools that better link knowledge with machine learning-based predictions from multi-scale, multi-modal, and longitudinal data; (2) innovative modeling approaches, such as causal inference techniques and hybrid modeling, which go beyond typical state-of-the-art machine learning; and (3) new computational modeling approaches that allow us to identify critical transitions in a patient's medical trajectory.

More speculatively, a broader understanding of human disease, incorporating findings from basic research and evolutionary studies, might help the creation of entirely new concepts for simulating human diseases and predicting optimal intervention points. Overall, the ambition of research towards personalized medicine should be to move from a system analysis perspective (such as in molecular biology) to a system control view that allows for the planning of optimal medical interventions at the right time and dose on an individualized basis. Novel computational modeling approaches that go beyond the current machine learning methodology may play an increasing role for that purpose.

In this context, it must be emphasized that no algorithm is meant to replace a physician. Rather, the idea is to provide them a tool at hand, which supports their decisions based on objective, data-driven criteria and the wealth of available biomedical knowledge.

Abbreviations

AI: Artificial Intelligence; CDM: core disease model; CLIA: Clinical Laboratory Improvement Amendments; EAM: environment adaption model; EMR: electronic medical record; FDA: Food and Drug Administration

Acknowledgements

We thank Colin Birkenbihl for providing Fig. 5. CB has been partially supported by the IMI project AETIONOMY (https://www.aetionomy.eu/en/vision.html) within the 7th Framework Programme of the European Union.

Funding

This work is an outcome of Dagstuhl seminar 17472 "Addressing the Computational Challenges of Personalized Medicine" (http://www.dagstuhl.de/de/programm/kalender/semhp/?semnr=17472), which was organized by authors HF, NB, and SM. We thank Schloss Dagstuhl for the support of our meeting. HF was partially supported by the IMI project AETIONOMY (https://www.aetionomy.eu/en/vision.html) within the 7th Framework Programme of the European Union. TMP was supported in part by the Intramural Research Program of the National Institutes of Health, National Library of Medicine. MS was partially supported by the Robert-Bosch Stiftung, Stuttgart, Germany, and the European Commission Horizon 2020 UPGx grant (668353).
The funders had no influence on the content of this manuscript.

Authors' contributions

All authors contributed to the content of this paper. HF guided the draft of the manuscript. All authors have read and approved the content.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]UCB Biosciences GmbH, Alfred-Nobel-Str. Str. 10, 40789 Monheim, Germany. [2]University of Luxembourg, 6 avenue du Swing, 4367 Belvaux, Luxembourg. [3]Department of Biosciences and Engineering, ETH Zurich, Mattenstr. 26, 4058 Basel, Switzerland. [4]University of Tübingen, WSI/ZBIT, Sand 14, 72076 Tübingen, Germany. [5]Department of Computer Science, University of Memphis, 2222 Dunn Hall, Memphis, TN 38152, USA. [6]Max-Planck-Institute for Informatics, 66123 Saarbrücken, Germany. [7]ETH Zurich, Seminar für Statistik, Rämistrasse 101, 8092 Zurich, Switzerland. [8]University of Leuven, ESAT, Kasteelpark Arenberg 10, 3001 Leuven, Belgium. [9]Harvard University, Science Center 400 Suite, Oxford Street, Cambridge, MA 02138-2901, USA. [10]National Center of Biotechnology Information, National Institute of Health, 8600 Rockville Pike, Bethesda, MD 20894-6075, USA. [11]Novartis Institutes for Biomedical Research, 4056 Basel, Switzerland. [12]Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, 160 College Street, Toronto, ON M5S 3E1, Canada. [13]RWTH Aachen, Joint Research Center for Computational Biomedicine, Pauwelsstrasse 19, 52074 Aachen, Germany. [14]Dr. Margarete Fischer-Bosch Institute of Clinical Pharmacology, Aucherbachstrasse 112, 70376 Stuttgart, Germany. [15]University of Regensburg, Institute of Functional Genomics, Am BioPark 9, 93053 Regensburg, Germany. [16]ETH Zurich, NEXUS Personalized Health Technol., Otto-Stern-Weg 7, 8093 Zurich, Switzerland. [17]Georgia Tech University, 801 Atlantic Drive, Atlanta, GA 30332-0280, USA. [18]Institute for Computer Science, University of Bonn, Endenicher Allee 19a, 53115 Bonn, Germany. [19]Pfizer, Worldwide Research and Development, Linkstraße 10, 10785 Berlin, Germany. [20]Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, SI-1000 Ljubljana, Slovenia. [21]University of Bonn, Bonn-Aachen International Center for IT, Endenicher Allee 19c, 53115 Bonn, Germany. [22]Max Planck Institute for Developmental Biology, Max-Planck-Ring 5, 72076 Tübingen, Germany. [23]Quantitative Biology Center, University of Tübingen, Auf der Morgenstelle 8, 72076 Tübingen, Germany. [24]Institute for Translational Bioinformatics, University Medical Center Tübingen, Sand 14, 72076 Tübingen, Germany. [25]University of Tübingen, Departments of Clinical Pharmacology and of Pharmacy and Biochemistry, Tübingen, Germany.

Received: 28 March 2018 Accepted: 9 July 2018
Published online: 27 August 2018

References

1. Sobradillo P, Pozo F, Agustí A. P4 medicine: the future around the corner. Arch Bronconeumol. 2011;47:35–40. https://doi.org/10.1016/j.arbres.2010.09.009.
2. Mathur S, Sutton J. Personalized medicine could transform healthcare. Biomed Rep. 2017;7:3–5. https://doi.org/10.3892/br.2017.922.
3. Vogenberg FR, Isaacson Barash C, Pursel M. Personalized medicine: part 1: evolution and development into theranostics. P T. 2010;35:560–76.
4. Hoffman MA, Williams MS. Electronic medical records and personalized medicine. Hum Genet. 2011;130:33–9. https://doi.org/10.1007/s00439-011-0992-y.
5. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012;13:395–405. https://doi.org/10.1038/nrg3208.
6. Lee CH, Yoon H-J. Medical big data: promise and challenges. Kidney Res Clin Pract. 2017;36:3–11. https://doi.org/10.23876/j.krcp.2017.36.1.3.
7. Lu J-J, Pan W, Hu Y-J, Wang Y-T. Multi-target drugs: the trend of drug research and development. PLoS One. 2012;7:e40262. https://doi.org/10.1371/journal.pone.0040262.
8. Vesell ES. Genetic and environmental factors causing variation in drug response. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 1991;247:241–57. https://doi.org/10.1016/0027-5107(91)90020-O.
9. van't Veer LJ, Dai H, van de Vijver MJ, He YD, AAM H, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415:530–6. https://doi.org/10.1038/415530a.
10. Cardoso F, van't Veer LJ, Bogaerts J, Slaets L, Viale G, Delaloge S, et al. 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. N Engl J Med. 2016;375:717–29. https://doi.org/10.1056/NEJMoa1602253.
11. Lengauer T, Sander O, Sierra S, Thielen A, Kaiser R. Bioinformatics prediction of HIV coreceptor usage. Nat Biotechnol. 2007;25:1407–10. https://doi.org/10.1038/nbt1371.
12. Lengauer T, Sing T. Bioinformatics-assisted anti-HIV therapy. Nat Rev Microbiol. 2006;4:790–7. https://doi.org/10.1038/nrmicro1477.
13. Büttner F, Winter S, Rausch S, Reustle A, Kruck S, Junker K, et al. Survival prediction of clear cell renal cell carcinoma based on gene expression similarity to the proximal tubule of the nephron. Eur Urol. 2015;68:1016–20. https://doi.org/10.1016/j.eururo.2015.05.045.
14. Lee J-G, Jun S, Cho Y-W, Lee H, Kim GB, Seo JB, et al. Deep learning in medical imaging: general overview. Korean J Radiol. 2017;18:570–84. https://doi.org/10.3348/kjr.2017.18.4.570.
15. Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. Annu Rev Biomed Eng. 2017;19:221–48. https://doi.org/10.1146/annurev-bioeng-071516-044442.
16. Djuric U, Zadeh G, Aldape K. Diamandis P. Precision histology: how deep learning is poised to revitalize histomorphology for personalized cancer care. npj Precision Onc. 2017;1:22. https://doi.org/10.1038/s41698-017-0022-1.
17. 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. https://doi.org/10.1038/srep26094.
18. Beaulieu-Jones BK, Orzechowski P, Moore JH. Mapping Patient Trajectories using Longitudinal Extraction and Deep Learning in the MIMIC-III Critical Care Database. Pac Symp Biocomput. 2018;23:123–32.
19. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. J Am Med Inform Assoc. 2017;24:361–70. https://doi.org/10.1093/jamia/ocw112.

20. 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.
21. Yu K-H, Zhang C, Berry GJ, Altman RB, Ré C, Rubin DL, et al. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nat Commun. 2016;7:12474. https://doi.org/10.1038/ncomms12474.
22. Fan J, Han F, Liu H. Challenges of big data analysis. Natl Sci Rev. 2014;1:293–314. https://doi.org/10.1093/nsr/nwt032.
23. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. Science. 2017;356:183–6. https://doi.org/10.1126/science.aal4230.
24. Mazzocchi F. Could Big Data be the end of theory in science? A few remarks on the epistemology of data-driven science. EMBO Rep. 2015;16:1250–5. https://doi.org/10.15252/embr.201541001.
25. Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set? Bioinformatics. 2005;21:171–8. https://doi.org/10.1093/bioinformatics/bth469.
26. Gönen M. Statistical aspects of gene signatures and molecular targets. Gastrointest Cancer Res. 2009;3(2 Suppl):S19–21.
27. Cun Y, Fröhlich HF. Prognostic gene signatures for patient stratification in breast cancer: accuracy, stability and interpretability of gene selection approaches using prior knowledge on protein-protein interactions. BMC Bioinformatics. 2012;13:69. https://doi.org/10.1186/1471-2105-13-69.
28. Cun Y, Fröhlich H. Biomarker gene signature discovery integrating network knowledge. Biology (Basel). 2012;1:5–17. https://doi.org/10.3390/biology1010005.
29. Chuang H-Y, Lee E, Liu Y-T, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol. 2007;3:140. https://doi.org/10.1038/msb4100180.
30. Lin W, Shi P, Feng R, Li H. Variable selection in regression with compositional covariates. Biometrika. 2014;101:785–97. https://doi.org/10.1093/biomet/asu031.
31. Altenbuchinger M, Schwarzfischer P, Rehberg T, Reinders J, Kohler CW, Gronwald W, et al. Molecular signatures that can be transferred across different omics platforms. Bioinformatics. 2017;33:i333–40. https://doi.org/10.1093/bioinformatics/btx241.
32. Rahmadi R, Groot P, Heins M, Knoop H, Heskes T. Causality on cross-sectional data: Stable specification search in constrained structural equation modeling. Appl Soft Comput. 2017;52:687–98. https://doi.org/10.1016/j.asoc.2016.10.003.
33. Maathuis MH, Colombo D, Kalisch M, Bühlmann P. Predicting causal effects in large-scale systems from observational data. Nat Methods. 2010;7:247–8. https://doi.org/10.1038/nmeth0410-247.
34. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005;102:15545–50. https://doi.org/10.1073/pnas.0506580102.
35. Bayerlová M, Jung K, Kramer F, Klemm F, Bleckmann A, Beißbarth T. Comparative study on gene set and pathway topology-based enrichment methods. BMC Bioinformatics. 2015;16:334. https://doi.org/10.1186/s12859-015-0751-5.
36. Fujita KA, Ostaszewski M, Matsuoka Y, Ghosh S, Glaab E, Trefois C, et al. Integrating pathways of Parkinson’s disease in a molecular interaction map. Mol Neurobiol. 2014;49:88–102. https://doi.org/10.1007/s12035-013-8489-4.
37. Funahashi A, Morohashi M, Kitano H, Tanimura N. CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. BIOSILICO. 2003;1:159–62. https://doi.org/10.1016/S1478-5382(03)02370-9.
38. Kutmon M, van Iersel MP, Bohler A, Kelder T, Nunes N, Pico AR, et al. PathVisio 3: an extendable pathway analysis toolbox. PLoS Comput Biol. 2015;11:e1004085. https://doi.org/10.1371/journal.pcbi.1004085.
39. Gawron P, Ostaszewski M, Satagopam V, Gebel S, Mazein A, Kuzma M, et al. MINERVA-a platform for visualization and curation of molecular interaction networks. npj Syst Biol Appl. 2016;2:16020. https://doi.org/10.1038/npjsba.2016.20.
40. Kuperstein I, Cohen DPA, Pook S, Viara E, Calzone L, Barillot E, et al. NaviCell: a web-based environment for navigation, curation and maintenance of large molecular interaction maps. BMC Syst Biol. 2013;7:100. https://doi.org/10.1186/1752-0509-7-100.
41. Hara S, Hayashi K. Making Tree Ensembles Interpretable: A Bayesian Model Selection Approach. In: Storkey A, Perez-Cruz F, editors. Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics: PMLR; 2018. p. 77–85.
42. Valdes G, Luna JM, Eaton E, Simone CB, Ungar LH, Solberg TD. Mediboost: a patient stratification tool for interpretable decision making in the era of precision medicine. Sci Rep. 2016;6:37854. https://doi.org/10.1038/srep37854.
43. JRR L, Kerridge I, Lipworth W. Use of Real-World Data for the Research, Development, and Evaluation of Oncology Precision Medicines. JCO Precis Oncol. 2017:1–11. https://doi.org/10.1200/PO.17.00157.
44. Breitenstein MK, Liu H, Maxwell KN, Pathak J, Zhang R. Electronic health record phenotypes for precision medicine: perspectives and caveats from treatment of breast cancer at a single institution. Clin Transl Sci. 2018;11:85–92. https://doi.org/10.1111/cts.12514.
45. Miksad RA, Abernethy AP. Harnessing the Power of Real-World Evidence (RWE): A Checklist to Ensure Regulatory-Grade Data Quality. Clin Pharmacol Ther. 2018;103:202–5. https://doi.org/10.1002/cpt.946.
46. Abernethy AP, Arunachalam A, Burke T, McKay C, Cao X, Sorg R, et al. Real-world first-line treatment and overall survival in non-small cell lung cancer without known EGFR mutations or ALK rearrangements in US community oncology setting. PLoS One. 2017;12:e0178420. https://doi.org/10.1371/journal.pone.0178420.
47. Kohavi R, Longbotham R. Online Controlled Experiments and A/B Testing. In: Sammut C, Webb GI, editors. Encyclopedia of Machine Learning and Data Mining. Boston: Springer; 2016. p. 1–8.
48. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a shared vision for cancer genomic data. N Engl J Med. 2016;375:1109–12. https://doi.org/10.1056/NEJMp1607591.
49. Ahmad A, Fröhlich H. Integrating Heterogeneous omics Data via Statistical Inference and Learning Techniques. Genomics Comput Biol. 2016;2:32. https://doi.org/10.18547/gcb.2016.vol2.iss1.e32.
50. Piening BD, Zhou W, Contrepois K, Röst H, Gu Urban GJ, Mishra T, et al. Integrative Personal Omics Profiles during Periods of Weight Gain and Loss. Cell Syst. 2018;6:157–170.e8. https://doi.org/10.1016/j.cels.2017.12.013.
51. Hinkson IV, Davidsen TM, Klemm JD, Kerlavage AR, Kibbe WA. A comprehensive infrastructure for big data in cancer research: accelerating cancer research and precision medicine. Front Cell Dev Biol. 2017;5:83. https://doi.org/10.3389/fcell.2017.00083.
52. Sagner M, McNeil A, Puska P, Auffray C, Price ND, Hood L, et al. The P4 Health Spectrum - A Predictive, Preventive, Personalized and Participatory Continuum for Promoting Healthspan. Prog Cardiovasc Dis. 2017;59:506–21. https://doi.org/10.1016/j.pcad.2016.08.002.
53. Beckmann JS, Lew D. Reconciling evidence-based medicine and precision medicine in the era of big data: challenges and opportunities. Genome Med. 2016;8:134. https://doi.org/10.1186/s13073-016-0388-7.
54. Deforche K, Camacho R, Van Laethem K, Lemey P, Rambaut A, Moreau Y, et al. Estimation of an in vivo fitness landscape experienced by HIV-1 under drug selective pressure useful for prediction of drug resistance evolution during treatment. Bioinformatics. 2008;24:34–41. https://doi.org/10.1093/bioinformatics/btm540.
55. Pearl J. Graphical models for probabilistic and causal reasoning. In: Smets P, editor. Quantified representation of uncertainty and imprecision. Dordrecht: Springer Netherlands; 1998. p. 367–89. https://doi.org/10.1007/978-94-017-1735-9_12.
56. Pearl J. Causality: Models, Reasoning and Inference. Cambridge: Cambridge University Press; 2000.
57. Spirtes P, Glymour C, Scheines R. Causation, Prediction and Search. Second edition. Cambridge: MIT Press. 2000.
58. Chickering DM. Learning equivalence classes of bayesian-network structures. Journal of Machine Learning Research. 2002;2:445–98.
59. Shimizu S, Hoyer PO, Hyvärinen A, Kerminen A. A linear non-Gaussian acyclic model for causal discovery. J Mach Learn Res. 2006;7:2003–30.
60. Heinze-Deml C, Maathuis MH, Meinshausen N. Causal Structure Learning. Annu Rev Stat Appl. 2017;5:371–391. https://doi.org/10.1146/annurev-statistics-031017-100630
61. Rathnam C, Lee S, Jiang X. An algorithm for direct causal learning of influences on patient outcomes. Artif Intell Med. 2017;75:1–15. https://doi.org/10.1016/j.artmed.2016.10.003.
62. Aliferis CF, Statnikov A, Tsamardinos I, Mani S, Koutsoukos XD. Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical Evaluation. Journal of Machine Learning Research. 2010;11 Jan:171–234.
63. Sun X, Janzing D, Schölkopf B, Fukumizu K. A kernel-based causal learning algorithm. In: Ghahramani Z, editor. Proceedings of the 24th international conference on Machine learning - ICML' ' '07. New York: ACM Press; 2007. p. 855–62. https://doi.org/10.1145/1273496.1273604.
64. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144:646–74. https://doi.org/10.1016/j.cell.2011.02.013.

65. Dingli D, Michor F. Successful therapy must eradicate cancer stem cells. Stem Cells. 2006;24:2603–10. https://doi.org/10.1634/stemcells.2006-0136.
66. von Stosch M, Oliveira R, Peres J, Feyo de Azevedo S. Hybrid semi-parametric modeling in process systems engineering: Past, present and future. Comput Chem Eng. 2014;60:86–101. https://doi.org/10.1016/j.compchemeng.2013.08.008.
67. Mogk G, Mrziglod T, Schuppert A. Application of hybrid models in chemical industry. In: European Symposium on Computer Aided Process Engineering-12, 35th European Symposium of the Working Party on Computer Aided Process Engineering. Elsevier; 2002. p. 931–936. https://doi.org/10.1016/S1570-7946(02)80183-3.
68. Psichogios DC, Ungar LH. A hybrid neural network-first principles approach to process modeling. AIChE J. 1992;38:1499–511. https://doi.org/10.1002/aic.690381003.
69. Fiedler B, Schuppert A. Local identification of scalar hybrid models with tree structure. IMA Journal of Applied Mathematics. 2008;73:449–76. https://doi.org/10.1093/imamat/hxn011.
70. Schuppert AA. Efficient reengineering of meso-scale topologies for functional networks in biomedical applications. JMathIndustry. 2011;1:6. https://doi.org/10.1186/2190-5983-1-6.
71. Balabanov S, Wilhelm T, Venz S, Keller G, Scharf C, Pospisil H, et al. Combination of a proteomics approach and reengineering of meso scale network models for prediction of mode-of-action for tyrosine kinase inhibitors. PLoS One. 2013;8:e53668. https://doi.org/10.1371/journal.pone.0053668.
72. Liu X, Chang X, Liu R, Yu X, Chen L, Aihara K. Quantifying critical states of complex diseases using single-sample dynamic network biomarkers. PLoS Comput Biol. 2017;13:e1005633. https://doi.org/10.1371/journal.pcbi.1005633.
73. Gluckman PD, Low FM, Buklijas T, Hanson MA, Beedle AS. How evolutionary principles improve the understanding of human health and disease. Evol Appl. 2011;4:249–63. https://doi.org/10.1111/j.1752-4571.2010.00164.x.
74. Jordan IK, Rogozin IB, Wolf YI, Koonin EV. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res. 2002;12:962–8. https://doi.org/10.1101/gr.87702.
75. Park S, Yang J-S, Kim J, Shin Y-E, Hwang J, Park J, et al. Evolutionary history of human disease genes reveals phenotypic connections and comorbidity among genetic diseases. Sci Rep. 2012;2:757. https://doi.org/10.1038/srep00757.
76. Hamey FK, Nestorowa S, Kinston SJ, Kent DG, Wilson NK, Göttgens B. Reconstructing blood stem cell regulatory network models from single-cell molecular profiles. Proc Natl Acad Sci USA. 2017;114:5822–9. https://doi.org/10.1073/pnas.1610609114.