The rocky road to personalized medicine: computational and statistical challenges

> “*On the road towards* in silico *humans with* in silico *diseases that are applicable to treatment prediction in personalized medicine, it is necessary to consider the intermediate steps where statistical models are built for the purposes of both prediction and systems level understanding, from the current and emerging data.*”

KEYWORDS: computational modeling ▪ data integration ▪ information retrieval ▪ machine learning ▪ ’omics data ▪ personalized medicine ▪ statistical prediction ▪ stratified medicine

Jukka Corander‡

Author for correspondence:
Department of Mathematics & Statistics, University of Helsinki, PO Box 68, 00014 Helsinki, Finland
and
Department of Mathematics, Åbo Akademi University, 20500 Åbo, Finland
jukka.corander@helsinki.fi
‡Author contributed equally

Tero Aittokallio‡

Institute for Molecular Medicine Finland (FIMM), University of Helsinki, 00014 Helsinki, Finland
and
Department of Mathematics, University of Turku, 20014 Turku, Finland
‡Author contributed equally

Samuli Ripatti‡

Institute for Molecular Medicine Finland (FIMM), University of Helsinki, 00014 Helsinki, Finland
and
Public Health Genomics Unit, National Institute for Health & Welfare, Helsinki, Finland
and
Wellcome Trust Sanger Institute, Hinxton, UK
‡Author contributed equally

Samuel Kaski‡

Helsinki Institute for Information Technology, Aalto University, 00076 Aalto, Finland
and
Helsinki Institute for Information Technology, University of Helsinki, Finland
‡Author contributed equally

The medical research community has over the past few years witnessed an intensive development of a broad array of technologies that has given a substantial push for the desideratum of a future personalized medicine, where prediction, prevention and treatment of illness are genuinely individualized [1–4]. The importance of this goal is demonstrated by the vigorous ongoing activities among academia, industry, patient group representatives and policy-makers to define an agreeable form of personalized medicine and to discuss its central aspects ranging from economical to ethical issues. A prime example of this activity is the Forward Look on Personalized Medicine initiated by the European Science Foundation, where representatives of a broad community are collaborating to establish a roadmap for European developments on personalized medicine for the next two decades [101].

A need for personalized medicine stems from several major factors, including the failure of the current research and development practices to develop effective therapies for an entire population of patients, the escalating costs of drug-development process under the prevailing standard of randomized clinical trials and the expected increasing disease burden due to the aging of populations throughout the world [5].

To ensure efficacy and safety of medical treatments, randomized clinical trials have, since several decades, been developed into the gold-standard approach to development of novel therapies and drugs. Such population-based studies in general fail to account for heterogeneity in the target population beyond crude stratifications, for example, with respect to age and gender. According to one predominant future scenario, in genuine personalized medicine such trials have given way to highly individualized therapy forms where predictions are based on combining detailed characteristics of the individual and the disease status with *in silico* models acting upon systems biology level of understanding of both the disease and humans as a whole [6–8]. For this scenario with its virtual patients to become reality in the foreseeable future, a plethora of challenges must be met, out of which we focus here on the computational and statistical issues that pave the road towards personalized medicine. In addition to these, there are numerous other crucial factors that will not be considered explicitly. For instance, to have a fully operational personalized medicine one must already have the truly individualized treatment options available, in other words, a large palette of molecules that match with response variation in different subpopulations of patients (e.g., genome subtypes) and disease (e.g., cancer subtypes).

Towards building systems biology-based *in silico* prediction models

The decline in the productivity of the pharmaceutical industry has greatly challenged the conventional approach of population-based therapeutic development, which relies on demonstrating favorable, yet averaged, treatment outcomes on a series of randomized controlled trials on large patient populations. Even though the shortcomings of such ‘one-size-fits-all’ treatment strategies are widely acknowledged, a major bottleneck hindering the development of novel and more selective treatment alternatives is the lack of systematic approaches that would be able to pinpoint the most effective therapeutic options or their combination for each patient. For instance, cancer subtypes may arise and develop from various genetic defects, and therefore, any given therapy often results in different treatment responses. Moreover, the underlying genetic heterogeneity results in alterations within multiple molecular pathways, which lead to various cancer phenotypes and make most tumors resistant to single agents. Therefore, systematic patient-based approaches to developing selective, targeted therapies for the spectrum of cancer subtypes are needed for more effective clinical outcomes [9,10]. However, the development of such personalized therapies remains to be problematic owing to a number of experimental, modeling and computational challenges. In the following commentary, we will elaborate these challenges in some detail and provide recommendations on how to address these issues using integrated experimental–computational approaches. It is not reasonable to assume that the transition from a silicon cell to a silicon human with its silicon diseases based on a detailed systems biology level understanding of the involved mechanisms, will all of a sudden, deliver the basis for genuine personalized medicine. Rather, the *in silico* prediction models and the related systems biology are expected to evolve gradually through feedback and refinement on the basis of statistical modeling of the predictions tested against real outcomes from individual patients.

> “*The decline in the productivity of the pharmaceutical industry has greatly challenged the conventional approach of population-based therapeutic development, which relies on demonstrating favorable, yet averaged, treatment outcomes on a series of randomized controlled trials on large patient populations.*”

Perhaps the most promising currently considered initial step towards genuine personalized medicine is stratified medicine [11], where the key task is to use multiple biomarkers jointly to identify subpopulations of patients who differ in terms of their disease traits or treatment outcomes. Since for an individual patient the link to the response via a population remains indirect and is only available for subjects that fit into a predefined fixed set of disease subpopulations, stratified medicine cannot be considered as truly personalized medicine. Results of stratified medicine can naturally be helpful, as known biomarkers can be used to identify the patient with a specific subpopulation when applicable, but fundamentally the fixed subpopulations can be interpreted as one of several possible model families applicable for arriving at the treatment predictions.

While reliable systems biology-based models remain absent, the treatment predictions need to be primarily learned from data, and it is impossible to learn to predict from data of a single patient before treatment outcomes have been measured. Moreover, robust and reliable predictions will remain elusive even when based on measured outcomes unless plenty of data are available. Hence, the statistical predictors need to combine information from observations that are sufficiently related. In stratified medicine, the additional strength would be borrowed from the subpopulation to which a patient is assigned based on the biomarker values, and consequently the prediction of a treatment outcome is assumed to be the same for all patients in the same subpopulation. An alternative approach, arguably more accurate for the individual patient, is to ask what would be the best patient-specific subpopulation to gain information from, and to learn both the subpopulation and the prediction at the same time. New statistical methods are needed for this task, and closely related problems are already being considered in the field of machine learning, for example, in the 2010 and 2011 workshops of the annual Neural Informations Processings Systems Foundation conference [102].

When approaching the ultimate level of personalized medicine, where each treatment group consists of a single individual, dense follow-up data will be an inevitable prerequisite for reliable inference and predictions. Such longitudinal experiments are based on experimental designs involving the baseline and several follow-up time points, for instance, before and after a particular disease status, intervention or development of resistance to a particular drug treatment. In this way, each individual is providing his or her own control measurement, thus enabling individual-level predictions of the disease progression or relapse, something that is not obtained on the basis of the cross-sectional case–control designs.

Integration of ’omics & cellular imaging datasets

Recent biotechnological advances in next-generation sequencing, tandem mass spectrometry, high-throughput screening, cellular imaging and so on, are enabling more indepth insights into the individual disease processes by generating millions of data points for each subject under analysis. However, the large-scale profiling experiments are notoriously prone to technical variability and instrument-specific biases; therefore, looking at any single data source alone will lead to limited and potentially biased views and predictions. For instance, the massive number of genetic variants found in genome-wide association studies or in exome/whole-genome sequencing makes it very challenging to distinguish between the variants truly associated with a disease and those originating from technical or disease-independent variability, leading to frequent false-positive and -negative findings. Moreover, the exponentially increasing number of potential interactions between the variants makes the pure experimental approach quickly unpowered, and translates into a need for integrated experimental–computational approaches that are scalable, robust and economical.

Efficient integration of complementary information sources from multiple levels, including tissue characteristics from cellular imaging, the genome, transcriptome, proteome, metabolome and interactome, can greatly facilitate the discovery of true causes and states of disease in specific subgroups of patients sharing a common genetic background. Recent activities in biocomputing demonstrate a vivid activity among these themes [12–18]. The issue of how to optimally combine multiple information sources in prediction tasks has also been recently studied extensively under the title of multiple kernel learning [19]. However, it is possible that such supervised data fusion methods may not help in the context of personalized medicine, given the severely limited number of observations per single feature. An alternative approach is to use unsupervised methods to identify what is shared between selected sets of data sources, ultimately basing the treatment effect predictions on the shared signals by assuming that they represent more relevant characteristics. Recent studies have shown how bioinformatics tools can be used to standardize and integrate measurements from complementary data sources with external knowledge bases, in order to reduce the effect of technology-specific noise and missing data and to boost statistical power [20]. In general, the road towards personalized medicine calls for harmonized statistical techniques for noise filtering applied both on individual datasets and at the metalevel of analysis.

Solid statistical framework for multidomain high-throughput patient-level data analysis

Unprecedented amount of new knowledge about the genetic architecture of common complex diseases has been accumulated over the past 5 years. Cataloging common genetic variation in large well phenotyped and population-based biobanks around the world has been the key factor for the success [103]. European cohorts and case–control samples have often been instrumental in these studies. Over the coming years, the biobank samples are going to provide a much more precise description of our genomic, metabolomics, proteomic and other high-throughput variation. However, owing to rapid technological advances, it is also likely that the resolution and precision of the data will vary within and between biobanks and cohorts. This is the current status in genomic data where generations of SNP chips and sequencing solutions have produced different sets of measured genetic variants. In disease genetics, stochastic imputation methods together with a control for uncertainty of the imputed variants have enabled pooling of data across genotyping platforms. Similar harmonization efforts combined with statistical and computational methods development will soon be needed for all other types of high-throughput data.

> *“When approaching the ultimate level of personalized medicine, where each treatment group consists of a single individual, dense follow-up data will be an inevitable prerequisite for reliable inference and predictions.”*

Although there are several European and international infrastructure development initiatives enhancing cooperation between biobanks and data resources, such as PHOEBE, P3G, BBMRI, ENGAGE and BIOSHARE, none of these efforts have focused on data-analysis challenges and specific data harmonization needs for personalized medicine. Key advances in pooling and analysis of structured data are needed to enable medicine to turn personal: solutions to enable large-scale pooling of well-harmonized data through federated systems across Europe, services enabling efficient analysis of multidomain high-throughput data, and statistical and computational methods development optimized for data federation and processing in cloud computing environments. The harmonization efforts need to cover the lifespan of individuals from pregnancy to late life and death in a prospective follow-up setting. As future personalized medicine is likely to be based on federated database structures, the analysis tools need to adapt to settings where the primary individual level data are not in one central repository.

Modeling, 'omics & cellular imaging datasets

On the road towards *in silico* humans with *in silico* diseases that are applicable to treatment prediction in personalized medicine, it is necessary to consider the intermediate steps where statistical models are built for the purposes of both prediction and systems level understanding, from the current and emerging data. The requirement of statistical modeling, as opposed to deterministic mathematical models, appears inevitable from the natural constraint that the uncertainty related to any particular outcome must be sensibly quantified to yield predictions that will be of real use in medical practices. Network modeling and machine learning -based approaches are also gaining popularity owing to their potential to effectively reduce the high-dimensional search spaces and to enable individual-level predictions for disease status or treatment outcomes.

From the data modeling point of view, the shift towards the 'large p, small n' setting, where the number of study variables (p) greatly exceeds the number of study samples (n), holds a great promise for medical research. The change also implies previously unseen unique modeling challenges. In particular, the traditional analytic modeling frameworks that were developed, typically under settings in which n exceeds p, will not be ideally suited for modern molecular data modeling, in which p greatly exceeds n [21]. Moreover, the nonlinear relationships between phenotypes and other patient characteristics, the modeling of which is fundamental for the development of personalized medicine strategies, pose modeling challenges beyond the reach of the classical linear and/or discrete models. Additional modeling challenges originate from the high complexity of the disease processes, including a multitude of genetic and environmental risk factors and interconnected pathways, which render detailed mathematical models prohibitively complicated. Also, the majority of the key factors and their interactions are currently unknown, making the predictions from the traditional bottom-up models unreliable. Therefore, in order to draw clinically relevant conclusions from the emerging high-dimensional datasets will likely require that the ongoing paradigm shifts in data generation and drug discovery will be accompanied by a similar paradigm shift in data modeling in the context of personalized medicine [21].

A particular modeling framework found useful in many biomedical applications is based on network graphs, which are simple to compute and interface when representing, integrating and mining high-dimensional experimental datasets. Hence, they provide an efficient framework for extracting and interpreting different types of relationships, such as those between genes and proteins, drugs and their cellular targets, or diseases [22–24]. Such coarser-level integration of 'omics and clinical datasets with network analyses and systems pharmacology offer holistic information on disease networks and drug action by considering chemical–target–disease relationships at network level, thus, enabling systematic computational strategies for personalized medicine [25,26]. Network modeling may help decipher how perturbations in the cellular networks lead to certain phenotypes, such as human diseases [27,28], and where in the disease networks one should target in order to inhibit the disease phenotypes, such as cancer progression [29,30]. Identification of disease-specific subnetworks, so-called modules, can reveal target pathways directly related to the disease process, and therefore, reduce both false positives and negatives caused by technical variability, secondary effects and/or genetic heterogeneity. Finally, by superimposing patient-based measurements into the global network models, such modular, or top-down, modeling approaches have the potential to enable individual-level disease predictions using network-level features, such as those based on subnetwork biomarkers of cancer metastasis [31,32].

> *"...the next generations of scientists who are acquiring training targeted towards solving personalized medicine-related problems will play a pivotal role on delivering the promises made today."*

In general, for predicting treatment outcomes for an individual patient we need a set of patient features potentially indicative of the outcome, measurements of the same features and known outcomes from a set of patients, and a model that learns the predictive mechanism from all available data. The grand challenge of machine learning in personalized medicine is to find the best approaches for this task that features multiple data sources, a very large numbers of variables p relative to the number of relevant observations n, and potentially very nonlinear predictions. Relevant machine learning genres include data integration or fusion to comply with the multiple input domains; recommender engine-type modeling for combining evidence from similar patients and evidence from the actual input data; dimensionality reduction and manifold estimation as preprocessing; various nonlinear regression methods; sparsity and regularization for complying with 'large p, small n'; multitask learning in case the data can be sensibly divided into subpopulations; and domain adaptation when moving to new types of patients.

In order to make progress towards the systems biology-based predictions, it is important to be able to understand what the treatment outcome predictions are based on even when applying machine learning tools. One possible option is to constrain the models to be better interpretable, for instance, by using rule-based models or expert systems. Since such constraints may potentially decrease the accuracy of the predictions, we would like to point at an alternative direction that provides additional methods for the analyst to visualize and interact with the predictions. As an example, one can consider a tool for retrieving from databases other patients for whom the predictions are similar, and based on similar combinations of the biomarkers as defined by the model. By studying the medical records and other characteristics of the retrieved patients, an analyst would be able to invoke his expertise or prior knowledge in interpreting the predictions. Tools for that task could be developed by extending methods recently applied to retrieving of earlier microarray experiments relevant to new measurement data [33,34], where the similarity can even be directly visualized, together with what the similarity is based on. Since building of the prediction models will necessarily be done in parallel within a large research community, it will also be necessary to consider tools for retrieval of relevant biological models for reuse in the same, related or alternative contexts [35].

Assuming that a prediction model is able to account for the differences between model organisms or cell lines and human subjects, retrieval of relevant samples may be applied to model organisms as well. Even treatment suggestions could then possibly be made given the comprehensive drug-response profile databases on cell lines [36], or based on treatments made on model organisms having similar biomarkers. The grand challenge is that a machine learning model then needs to be able to correctly translate predictions between the model organisms and humans, which has proven to be extremely difficult in general [37].

Given that several sources of uncertain evidence will be needed in making the predictions, it may be difficult to justify making treatment choices based on predictions of the abstract models, no matter how well interpretable they are. The use of the models can therefore perhaps be most easily justified in drug repositioning, when using already approved drugs [38].

Conclusion

Personalized medicine is one of the grandest challenges that the medical research community has ever faced. When successful, it will have huge implications for human societies. The obstacles on the road towards personalized medicine require entirely novel thinking in terms of acquisition and use of data to develop useful and reliable treatment outcome predictions for individual patients. In particular, the unprecedented scale of the challenge calls for a tighter integration of the efforts of individual members of the research community through development of strategies that enable faster and more manageable sharing and combination of information and knowledge. Also, the next generations of scientists who are acquiring training targeted towards solving personalized medicine-related problems will play a pivotal role on delivering the promises made today.

Future perspective

Currently, the amount and complexity of biological and medical data is increasing more rapidly than the computational power, storage facilities and modeling techniques improve. This poses an enormous intellectual challenge to the information and communications technology sector. Progress to a future personalized medicine is also expected to vary largely over the different disease types, such as cardiovascular diseases, cancer, and CNS-related and rare diseases. Different disease types will each require their own 'subparadigms' of personalized medicine. We anticipate that the developments will be fastest for cancer where deeply stratified medicine should have reached a widely established role within a timespan of 5–10 years. Simultaneously, we expect that the integration of multiple high-throughput data sources and community efforts will yield a significant leap towards a systems biology level understanding of causes for a wide variety of diseases. Data harmonization on national, European and international levels, computational efficiency and rigorous statistical inference are key elements for success in the coming era of personalized medicine.

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

References

1 Auffray C, Caulfield T, Khoury MJ, Lupski JR, Schwab M, Veenstra T. Genome medicine: past, present and future. *Genome Med.* 3(1), 6 (2011).

2 Auffray C, Chen Z, Hood L. Systems medicine: the future of medical genomics and healthcare. *Genome Med.* 1(1), 2 (2009).

3 Hamburg MA, Collins FS. The path to personalized medicine. *N. Engl. J. Med.* 363(4), 301–304 (2010).

4 Offit K. Personalized medicine: new genomics, old lessons. *Hum. Genet.* 130(1), 3–14 (2011).

5 Sullivan R, Peppercorn J, Sikora K *et al.* Delivering affordable cancer care in high-income countries. *Lancet Oncol.* 12(10), 933–980 (2011).

6 Westerhoff HV, Kolodkin A, Conradie R *et al.* Systems biology towards life *in silico*: mathematics of the control of living cells. *J. Math. Biol.* 58(1–2), 7–34 (2009).

7 Westerhoff HV, Verma M, Bruggeman FJ *et al. From Silicon Cell to Silicon Human BetaSys.* Booβ-Bavnbek B, Klösgen B, Larsen J, Pociot F, Renström E (Eds). Springer, NY, USA, 437–458 (2011).

8 Bruggeman FJ, Westerhoff HV. Approaches to biosimulation of cellular processes. *J. Biol. Phys.* 32(3–4), 273–288 (2006).

9 Deisboeck TS. Personalizing medicine: a systems biology perspective. *Mol. Syst. Biol.* 5, 249 (2009).

10 Schreiber SL, Shamji AF, Clemons PA *et al.* Towards patient-based cancer therapeutics. *Nat. Biotechnol.* 28(9), 904–906 (2010).

11 Trusheim MR, Berndt ER, Douglas FL. Stratified medicine: strategic and economic implications of combining drugs and clinical biomarkers. *Nat. Rev. Drug Discov.* 6(4), 287–293 (2007).

12 Lundin J, Lundin M, Isola J, Joensuu H. A web-based system for individualised survival estimation in breast cancer. *BMJ* 326(7379), 29–29 (2003).

13 Lundin M, Szymas J, Linder E *et al.* A European network for virtual microscopy-design, implementation and evaluation of performance. *Virchows Arch.* 454(4), 421–429 (2009).

14 Marttinen P, Myllykangas S, Corander J. Bayesian clustering and feature selection for cancer tissue samples. *BMC Bioinformatics* 10, 90 (2009).

15 Curtis RE, Yin J, Kinnaird P, Xing EP. Finding genome-transcriptome-phenome associations with structured association mapping and visualization in genamap. *Pac. Symp. Biocomput.* 17, 327–328 (2012).

16 Karczewski KJ, Tirrell RP, Cordero P *et al.* Interpretome: a freely available, modular, and secure personal genome interpretation engine. *Pac. Symp. Biocomput.* 17, 339–350 (2012).

17 Pal R, Berlow N. A kinase inhibition map approach for tumor sensitivity prediction and combination therapy design for targeted drugs. *Pac. Symp. Biocomput.* 17, 351–362 (2012).

18 Warde-Farley D, Brudno M, Morris Q, Goldenberg A. Mixture model for sub-phenotyping in GWAS. *Pac. Symp. Biocomput.* 17, 363–374 (2012).

19 Gonen M, Alpaydin E. Multiple kernel learning algorithms. *J. Machine Learning Res.* 12, 2211–2268 (2011).

20 Ideker T, Dutkowski J, Hood L. Boosting signal-to-noise in complex biology: prior knowledge is power. *Cell* 144(6), 860–863 (2011).

21 Marko NF, Weil RJ. Mathematical modeling of molecular data in translational medicine: theoretical considerations. *Sci. Transl. Med.* 2(56), 56rv4 (2010).

22 Aittokallio T, Schwikowski B. Graph-based methods for analysing networks in cell biology. *Brief. Bioinform.* 7(3), 243–255 (2006).

23 Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* 12(1), 56–68 (2011).

24 Vidal M, Cusick ME, Barabasi AL. Interactome networks and human disease. *Cell* 144(6), 986–998 (2011).

25 Hopkins AL. Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.* 4(11), 682–690 (2008).

26 Wist AD, Berger SI, Iyengar R. Systems pharmacology and genome medicine: a future perspective. *Genome Med.* 1(1), 11 (2009).

27 Del Sol A, Balling R, Hood L, Galas D. Diseases as network perturbations. *Curr. Opin. Biotechnol.* 21(4), 566–571 (2010).

28 Schadt EE. Molecular networks as sensors and drivers of common human diseases. *Nature* 461(7261), 218–223 (2009).

29 Pe'er D, Hacohen N. Principles and strategies for developing network models in cancer. *Cell* 144(6), 864–873 (2011).

30 Schadt EE, Friend SH, Shaywitz DA. A network view of disease and compound screening. *Nat. Rev. Drug Discov.* 8(4), 286–295 (2009).

31 Chuang HY, Hofree M, Ideker T. A decade of systems biology. *Ann. Rev. Cell Dev. Biol.* 26, 721–744 (2010).

32 Pujol A, Mosca R, Farres J, Aloy P. Unveiling the role of network and systems biology in drug discovery. *Trends Pharmacol. Sci.* 31(3), 115–123 (2010).

33 Caldas J, Gehlenborg N, Faisal A, Brazma A, Kaski S. Probabilistic retrieval and visualization of biologically relevant microarray experiments. *Bioinformatics* 25(12), I145–I153 (2009).

34 Le HS, Oltvai ZN, Bar-Joseph Z. Cross-species queries of large gene expression databases. *Bioinformatics* 26(19), 2416–2423 (2010).

35 Köhn D, Maus C, Henkel R *et al.* Towards enhanced retrieval of biological models through annotation-based ranking. In: *Data Integration in the Life Sciences.* Paton NW, Missier P, Hedeler C (Eds). Springer, Berlin/Heidelberg, Germany, 204–219 (2009).

36 Lamb J, Crawford ED, Peck D *et al.* The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313(5795), 1929–1935 (2006).

37 Sysi-Aho M, Ermolov A, Gopalacharyulu PV *et al.* Metabolic regulation in progression to autoimmune diabetes. *PLoS Comput. Biol.* 7(10), e1002257 (2011).

38 Lussier YA, Chen JL. The emergence of genome-based drug repositioning. *Sci. Transl. Med.* 3(96), 96ps35 (2011).

Websites

101 Personalised Medicine for the European citizen – towards more precise medicine for the diagnosis, treatment and prevention of disease.
www.esf.org/iPM

102 Neural Informations Processings Systems Foundation.
http://nips.cc

103 A catalog of published genome-wide association studies.
www.genome.gov/gwastudies