Big data opens a window onto wellness

Atul J Butte

Longitudinal multi-omics data, clinical tests and biomarker analyses across a large cohort lay the groundwork for understanding the transition from wellness to disease.

It is difficult to make predictions, especially about the future [Anon.].

The idea of predicting which future diseases might afflict an individual is decades old. Nowadays, clinicians have access to a plethora of diagnostics, including biomarkers, biochemical assays and explanatory genetic associations. For example, the presence of *BRCA1* variants can be used to predict individual risk for breast and ovarian cancers, and the Framingham Risk Score can be applied to predict 10-year risk for coronary heart disease[1]. Advocates of 'systems medicine' claim that integration of analytical technologies with multi-omics, including genome sequences, could improve predictions of individual health trajectories for many diseases. Reporting in this issue, Price *et al.*[2] describe how 108 apparently healthy individuals were studied over 9 months with a greater level of resolution of multiple measurements than has previously been attempted on a population of this size. It is conceivable that data generated in this and other studies could eventually inform efforts to prevent disease.

Preventive medicine is not new. The US Preventive Services Task Force (USPTF), for example, lists 98 guidance documents for screening, counseling and preventive medications (https://www.uspreventiveservicestaskforce.org/BrowseRec/Index/browse-recommendations), but this guidance applies to the entire population rather than to individuals. Importantly, the USPTF does not conduct R&D into preventive strategies, but there are likely to be useful prognostic biomarkers and preventive measures waiting to be discovered, or validated, for clinical use. To improve prevention medicine, researchers have begun to recruit populations of healthy individuals to try and better define health and disease, to understand the health-to-disease transition, and to keep healthy those who are well. Typically, these studies evaluate high-density measurements, like exome- or genome-wide sequencing, blood RNA profiling, circulating DNA, cytokine or other protein measurements, cell counts and identification, wearable measurements and more, over time[3].

Price *et al.*[2] recruited a cohort of 108 individuals who were monitored for 9 months in a pilot study that the authors named the "Pioneer 100 Wellness Project." Various samples, including blood and feces, were collected at three time points during the 9 months. The number of measurements made using these samples was immense: whole genome sequencing, a battery of clinical tests, metabolomes, quantitation of plasma proteins and peptides, and identification of the gut microbiome. In addition, Price *et al.*[2] report the consistency of their measurements, including clinical laboratory tests. Correlations were run on duplicate measurements across vendors, and it is eye-opening that many standard clinical blood tests do not correlate well across technologies. In addition, home measurements and daily activity tracking were set up, but sadly the compliance rates for these were very low. The authors used systems biology computational methods to produce data clouds for each person and integrated

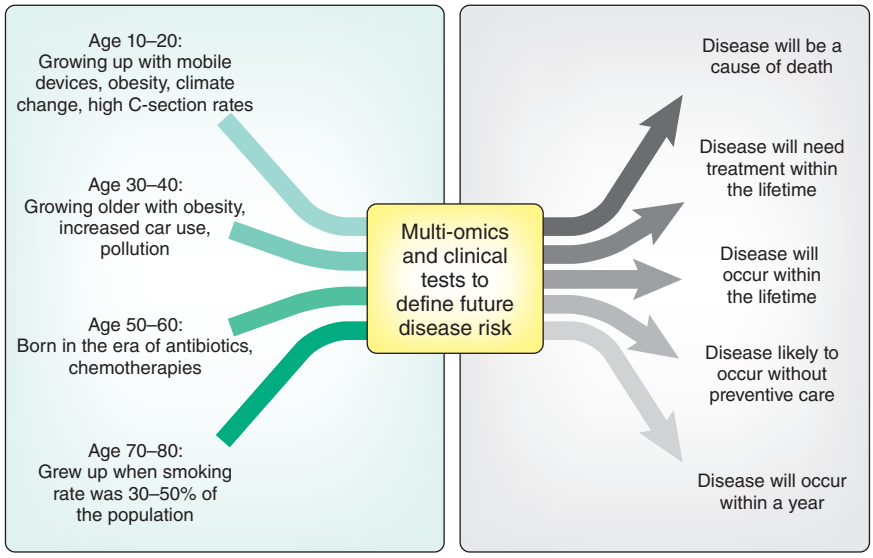

Figure 1 An individual undergoes multi-modality molecular testing. Every individual that is tested has a different pre-test probability of disease, which depends on when and where s/he grew up, environmental exposures, tobacco, alcohol and drug use, nutrition over time and much more (left-hand panel). If a complex set of tests (including germline DNA sequencing) indicates the risk for a particular disease, what exactly does this mean? Does this mean a disease might appear in the next year, or sometime during a lifetime (right-hand panel)? Should an individual be concerned about such risks? Modeling of additional variables pertinent to pre-test and post-test probabilities is needed now, so that individuals and medical professionals can derive specific utility from test data.

Atul J. Butte is at the Institute for Computational Health Sciences, University of California, San Francisco, USA.
e-mail: atul.butte@ucsf.edu

these to discover potentially novel molecular markers for pre-disease physiological conditions, such as obesity.

This was not a simple observational study. Ten coaching sessions were offered during the course of the study, in an attempt to improve individual health, but exactly how these measured data sets and risk predictions were translated into action points for coaches remains to be documented in depth in future publications. In this manuscript, the authors report that predictions of risk in four main categories (cardiovascular, diabetes, inflammation, and nutrition) were converted into behavioral and dietary advice that was in line with current medical guidance, but was tailored for each individual's lifestyle.

This latest report[2] complements recent studies that have examined a wealth of parameters in large human cohorts. These include the fitness of 43 individuals, as revealed using health and environmental trackers, which outputted 250,000 daily measurements[4]; analysis of a cohort of 10,000 people using genome sequencing[5]; and exome sequencing of a cohort of 50,000 patients cared for in a conventional medical context[6]. Other big data efforts include the collection of baseline measurements (including imaging) for 500,000 individuals by the UK Biobank[7], and the US Precision Medicine Initiative's All of US Research Program, which will gather molecular, lifestyle and environmental measurements on a million or more people (http://allofus.nih.gov/). Classic long-running efforts, such as the US Centers for Disease Control and Prevention's National Health and Nutrition Examination Survey and the Framingham Heart Study, which now includes microbiome analyses, still hold value for preventive medicine.

What lessons can we learn from these longitudinal big data studies?

One of the individuals in Price *et al.*[2] was found to have high ferritin levels in analytic tests. Genome sequencing confirmed a diagnosis of hereditary hemochromatosis, so we learn that genomes can provide useful medical information to healthy individuals. Asymptomatic individuals were unexpectedly discovered to harbor genetic risk factors for treatable conditions, including familial hypercholesterolemia and breast cancer in an unrelated study[6]. Are we nearing an inflexion point, when DNA sequencing could be a screening tool in early adulthood? Perhaps, but many questions remain unanswered. Whole genomes, or exomes, or just a few genes, for example? Multiple approaches exist for combining risk alleles[8], but a comprehensive comparison of these methods hasn't been made. Even more pressing is an understanding of what we mean by predicting that an individual is at risk for a disease, especially when that individual comes to the test with very different pre-test probabilities of disease[9] (**Fig. 1**).

It is well-known that health typically improves with more frequent medical attention, consistent with the findings of Price *et al.*[2]. Without a comparison arm, it is not possible to conclude that the health improvements they report are due to the big data analyses. Going forward, it is important to compare with control individuals receiving conventional preventive medicine. Not only should future studies compare results with those of conventional care, participants in 'intervention arms' should also receive conventional medical care. For example, risk scores that are standard of care must also be applied to these cohorts. How much of the guidance used in multi-omics-guided coaching might have been revealed by a high Framingham Risk Score, for example?

With regards to standard preventive medicine, Price *et al.*[2] note that 47 individuals had an out-of-range fasting glucose at baseline, which is indicative of pre-diabetes. The USPTF recommends that overweight adults aged 40 to 70 years of age should be screened for abnormal blood glucose (https://www.uspreventiveservicestaskforce.org/Page/Document/RecommendationStatementFinal/screening-for-abnormal-blood-glucose-and-type-2-diabetes). Did all participants receive routine care from their health providers? Results of big data studies might be more generalizable if they were benchmarked against standard medical care.

With less than one year of measurements, outcomes might be subject to seasonal confounders. The most significantly improved clinical variable in Price *et al.*[2] was vitamin D concentration. Price *et al.*[2] started their study in April when vitamin D levels in Americans are at their lowest[10]. The authors accounted for seasonal sun exposure in their analyses, but ideally studies should run for more than a year, to compensate for any unaccounted for seasonal effects.

Provision of the entire data set by Price *et al.*,[2] will be a boon for researchers, but getting one's own data cloud will not come cheap. Arivale (Seattle) offers wellness testing for $3,499 (https://www.arivale.com/pricing/). Human Longevity (San Diego), a competitor, offers services for $25,000–50,000 (depending on the panel of tests)[11]. This puts the price tag for participation in these cohorts beyond most people. And neither of these services can replace conventional medical care. Whether these commercial endeavors can attract a diverse enough population to make the research findings generalizable remains to be seen.

We are reaching a point when results from studies like this one may be used to augment standard care. Will health systems be able to process and interpret big data? And how much evidence is needed to justify the costs of comprehensive molecular testing?

Deciding whether to expand preventive services will depend on costs and demonstrated efficacy with strictly defined outcomes for a diverse general population. We are clearly still at the very beginning of understanding how precision medicine might benefit public health.

ACKNOWLEDGMENTS

A.J.B. is supported by the National Institutes of Health under award number 1OT2OD024611. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health.

COMPETING FINANCIAL INTERESTS

The author declares competing financial interests: details are available in the online version of the paper.

1. Wilson, P.W. *et al. Circulation* **97**, 1837–1847 (1998).
2. Price, N.D. *et al. Nat. Biotechnol.* **35**, 747–756 (2017).
3. Sheridan, C. *Nat. Biotechnol.* **33**, 887–889 (2015).
4. Li, X. *et al. PLoS Biol.* **15**, e2001402 (2017).
5. Telenti, A. *et al. Proc. Natl. Acad. Sci. USA* **113**, 11901–11906 (2016).
6. Dewey, F.E. *et al. Science* **354**, aaf6814 (2016).
7. Sudlow, C. *et al. PLoS Med.* **12**, e1001779 (2015).
8. Morgan, A.A., Chen, R. & Butte, A.J. *Genome Med.* **2**, 30 (2010).
9. Butte, A.J. *Proc. Natl. Acad. Sci. USA* **113**, 12344–12346 (2016).
10. Kasahara, A.K., Singh, R.J. & Noymer, A. *PLoS One* **8**, e65785 (2013).
11. Taylor, N.P. *Fierce Biotech* http://www.fiercebiotech.com/genomics/venter-s-human-longevity-starts-50-000-health-testing-service (2015).