The coming age of data-driven medicine: translational bioinformatics' next frontier

Nigam H Shah and Jessica D Tenenbaum

J Am Med Inform Assoc 2012 19: e2-e4
doi: 10.1136/amiajnl-2012-000969

Updated information and services can be found at:
http://jamia.bmj.com/content/19/e1/e2.full.html

These include:

References
This article cites 43 articles, 16 of which can be accessed free at:
http://jamia.bmj.com/content/19/e1/e2.full.html#ref-list-1

Article cited in:
http://jamia.bmj.com/content/19/e1/e2.full.html#related-urls

Open Access
This is an open-access article distributed under the terms of the Creative Commons Attribution Non-commercial License, which permits use, distribution, and reproduction in any medium, provided the original work is properly cited, the use is non commercial and is otherwise in compliance with the license. See:
http://creativecommons.org/licenses/by-nc/2.0/ and
http://creativecommons.org/licenses/by-nc/2.0/legalcode.

Email alerting service
Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Topic Collections
Articles on similar topics can be found in the following collections
Open access (116 articles)

Notes

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/
The coming age of data-driven medicine: translational bioinformatics’ next frontier

Nigam H Shah,1 Jessica D Tenenbaum2

Last year, in 2011, we argued that biomedical informatics stands ready to revolutionize human health and healthcare using large-scale measurements on a large number of individuals.1 We anticipated that, with the coming changes in the amount and diversity of datasets, data-centric approaches that compute on massive amounts of data (often called ‘Big Data’2 3) to discover patterns and to make clinically relevant predictions would be increasingly common in translational bioinformatics.

Given these trends, we programmed the 2012 Summit on Translational Bioinformatics to focus on research that takes us from base pairs to the bedside,4 with a particular emphasis on clinical implications of mining massive datasets, and bridging the latest multimodal measurement technologies with the large amounts of electronic healthcare data that are increasingly available.

The coming year did turn out to be the year of Big Data for the Summit, with multiple submissions on managing and interpreting large datasets (figure 1). Among the 35 full paper submissions to the Summit, four stood out for their association study (GWAS) data, Russu et al12 introduced a novel Bayesian model search algorithm, binary outcome stochastic search, for model selection when the number of predictors (eg, SNPs) far exceeds the number of observations. They propose an innovative stochastic model search technique where the relationship between the observed responses and the available predictors is described by a latent variable model with a probit link. They compare binary outcome stochastic search with three established methods (stepwise regression, logistic lasso, and elastic net) in a simulated study and in two real world studies to demonstrate higher precision (while preserving recall) in identifying SNPs associated with the observed outcome than the one obtained from established methods.

Morgan et al13 recipient of the Marco Ramoni Best Paper Award, constructed genomic disease risk summaries for 55 common diseases using reported gene—disease associations in the research literature. They constructed risk profiles based on the SNPs as well as on 187 whole-genome sequences and show that risk predictions derived from sequencing differ substantially from those obtained from the SNPs for several different non-monogenic diseases. When a large fraction of associated variants for a given disease is not covered by the genotyping array, the overall risk predictions can vary dramatically—by as much as a factor of 20 times in some instances.

Beyond this year’s conference papers, in the larger informatics community, researchers have demonstrated that GWAS can now be performed by leveraging large amounts of electronic medical record (EMR) data. For example, Kho et al showed that, by using commonly available data from five different EMRs, it is possible to accurately identify type 2 diabetes cases and controls for genetic study across multiple institutions.14 In addition, genomic sequencing has moved out of the research realm and established itself in the clinic. For example, at the Medical College of Wisconsin, Dr Howard Jacob’s team used genome sequencing to identify a novel causal mutation that led to successful treatment of a 6-year-old boy with an extreme form of inflammatory bowel disease.15 16

Currently, the discussion of Big Data in translational informatics often connotes next-generation sequencing data.3 17 18 However, this is beginning to change: in 2011, the use of large public datasets of various kinds increased dramatically. The research activity around data mining for predicting adverse drug events (ADEs) using public data is an excellent example.19 Drug safety surveillance is currently based on spontaneous reporting systems, which contain reports of suspected ADEs seen in clinical practice. In the USA, the primary database for such reports is the Adverse Event Reporting System (AERS) database at the Food and Drug Administration. This resource has been successfully mined using ‘disproportionality measures’, which quantify the magnitude of difference between observed and expected rates of particular drug—ADE pairs.20 21

Given the amount of data available in AERS,22 researchers are developing methods for detecting new or latent multi-drug adverse events. Examples include using side effect profiles from AERS’ reports to infer the presence of unreported adverse events,23—25 and...
creating a network of known drug—ADE relationships to predict as yet unknown ADEs before they are found in post-market evidence.36

Going beyond reported adverse events and making use of molecular level data, Pouliot et al37 generated logistic regression models to correlate and predict post-marketing ADEs based on screening data from PubChem, a public database of chemical structures of small organic molecules along with information about their biological activities. In a related effort, Vilari et al38 devised a way to enhance existing, data-mining algorithms with chemical information using molecular fingerprints—which represent molecules through a bit vector that codifies the existence of particular structural features or functional groups—to enhance ADE signals generated from adverse event reports. There have been increasing efforts to use other data sources, such as EMRs, for the purpose of detecting ADEs39–41 and to discover multi-drug ADEs.32

Researchers have also used billing and claims data for active drug safety surveillance33–35 and applied literature mining for drug safety.36 Recently, Chee et al37 explored the use of online health forums as a source of data to identify drugs for further scrutiny. They aggregate individuals’ opinions of drugs in roughly 12 million personal health messages using natural language processing and are able to identify drugs withdrawn from the market based on messages discussing them before their removal.38

Looking ahead, we believe that Big Data in biomedical informatics will be far more than genome sequence data.36–40 We argue that Big Data must be considered in a comprehensive manner, including both large amounts of ‘molecular measurements’ on a person (eg, sequencing) and small amounts of ‘routine measurements’ on a large number of people (eg, clinical notes, laboratory measurements, claims data and adverse event reports). In contrast with the buzz around genomic-data-in-the-clinic or adverse event predictions, consider the example by Frankovich et al.41 When the existing literature and a survey of colleagues was insufficient to guide the clinical care of a patient, Frankovich et al applied trend analysis to the EMR data from 98 patients to ‘learn’ a data-driven guideline on how to provide care for a 13-year-old girl with systemic lupus erythematosus.41 Such data-centric approaches are particularly useful when derivation of a formal guideline is not feasible from a practical standpoint.

It is tantalizing to imagine how scientific inquiry would be performed differently if we collect and share access to lots of data—both genomic and ‘routine’. How will the kinds of questions we ask change when we cross a certain data threshold?42 43 For example, researchers at Carnegie Mellon University built a scene completion tool by scraping millions of other images on the web from public sources. After the system accumulated a corpus of millions of photos, completed scenes were indistinguishable to the naked eye. The case for Big Data analytics has already won over the legal domain in at least one application, replacing armies of lawyers with computer algorithms designed for ‘e-discovery’—that is, retrieval of relevant materials for a legal case.44 Even the liberal arts are embracing Big Data: capitalizing on Google’s efforts to digitize books, researchers in the humanities are blazing new trails in ‘culturomics’ by examining language based on the analysis of word combinations occurring in millions of digitized books through time.45

In 2013, we will have the sixth Summit on Translational Bioinformatics and the third year of the AMIA Joint Summits on Translational Science. Translational research has become integral to the biomedical research enterprise, as evidenced by the creation of a National Center for Advancing Translational Science at the NIH. The Joint Summits continue to be a venue to facilitate dramatic changes that are underway to deliver quality, personalized healthcare in the USA without increasing spending at a rate exceeding the growth of the GDP.46 Reflecting this priority, the 2013 TBI Summit will have new tracks that will showcase the ways in which the translational sciences are having a significant impact on the way clinical care, biomedical research, and drug discovery are performed.

We believe that the time is ripe for medicine to embrace Big Data, to usher in the age of data-driven medicine—and to truly enable proactive, predictive, preventive, participatory, and patient-centered health.47 Data-driven medicine will enable the discovery of new treatment options based on the multi-model molecular measurements on patients and learning from the trends hidden among the diagnoses, prescriptions, and discharge summaries of millions of patient encounters logged by clinical practitioners.48 49

The increasing synergy between the Translational Bioinformatics Summit and the Clinical Research Informatics Summit is an indication of this impending convergence. This is an exciting time when medicine begins utilizing massive amounts of data to discover patterns and trends and to make predictions in a manner that is a mainstay of web-scale computing.42

Funding NHS is funded by the US National Institute of Health Roadmap (U54 HG004028 and U54 LM008748). JDT is funded by a Clinical and Translational Science Award (UL1 RR024128) and a gift from David H Murdock.

Competing interests None.

Provenance and peer review Commissioned; internally peer reviewed.

Accepted 26 March 2012

Figure 1 A tag cloud generated from the title and abstracts of the submissions made to the AMIA Translational Bioinformatics Summit 2012. The more frequently used the words are, the larger they appear. ‘Data’ was the most commonly mentioned word across all submissions for 2012.
REFERENCES


