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

Nigam H Shah,[1] Jessica D Tenenbaum[2]

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 innovation, and hence the authors were invited to expand the work for this special issue of *JAMIA*—adding to the growing presence of translational bioinformatics in the journal.[5–9]

Liu *et al*[10] demonstrated how the ability to predict adverse drug reactions can be increased by integrating chemical, biological, and phenotypic properties of drugs. They demonstrated that prediction accuracy increased from 0.9054 (when only chemical structures were used) to 0.9524 (when chemical structures along with biological and phenotypic features were used). They conclude that data fusion approaches are promising for large-scale adverse drug reaction predictions in both preclinical and post-marketing phases.

Bhavnani *et al*[11] assert that existing methods to analyze ancestral informative single-nucleotide polymorphisms (SNPs) (ie, SNPs that have large differences in genotype frequencies between two or more ancestral populations) identify a parsimonious set of SNPs that can identify distinct population clusters. However, existing methods do not directly visualize which clusters of subjects are related to which clusters of SNPs, or allow visualization of the genotypes that determine the cluster memberships. In an attempt to reveal such hidden relationships, they used three bipartite analytical representations (a bipartite network, a heat map with dendrograms, and a Circos ideogram) to simultaneously visualize clusters of subjects, SNPs, and the attributes that cause them to cluster.

Seeking to maximize the utility of the abundance of available genome-wide association study (GWAS) data, Russu *et al*[12] 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 al*,[13] 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

[1]Stanford Center for Biomedical Informatics Research, Stanford University School of Medicine, Stanford, California, USA; [2]Duke Translational Medicine Institute, Duke University, Durham, North Carolina, USA

Correspondence to Dr Nigam H Shah, Stanford University School of Medicine, 1265 Welch Road, Room X-229, Stanford, CA 94305, USA; nigam@stanford.edu

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.

creating a network of known drug–ADE relationships to predict as yet unknown ADEs before they are found in post-market evidence.[26]

Going beyond reported adverse events and making use of molecular level data, Pouliot *et al*[27] 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, Vilar *et al*[28] 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 ADEs[29–31] and to discover multi-drug ADEs.[32] Researchers have also used billing and claims data for active drug safety surveillance[33–35] and applied literature mining for drug safety.[36] Recently, Chee *et al*[37] 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.

Looking ahead, we believe that Big Data in biomedical informatics will be far more than genome sequence data.[38–40] 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

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

REFERENCES

1. **Butte AJ,** Shah NH. Computationally translating molecular discoveries into tools for medicine: translational bioinformatics articles now featured in JAMIA. *J Am Med Inform Assoc* 2011;**18**:352–3.
2. **Trelles O,** Prins P, Snir M, *et al*. Big data, but are we ready? *Nat Rev Genet* 2011;**12**:224.
3. **Schadt EE,** Linderman MD, Sorenson J, *et al*. Cloud and heterogeneous computing solutions exist today for the emerging big data problems in biology. *Nat Rev Genet* 2011;**12**:224.
4. **Green ED,** Guyer MS; National Human Genome Research Institute. Charting a course for genomic medicine from base pairs to bedside. *Nature* 2011;**470**:204–13.
5. **Berg JM.** National centers for biomedical computing: from the BISTI report to the future. *J Am Med Inform Assoc* 2012;**19**:151–2.
6. **Ohno-Machado L.** Informatics research to enable clinically relevant, personalized genomic medicine. *J Am Med Inform Assoc* 2012;**19**:149–50.
7. **Floratos A,** Honig B, Pe'er D, *et al*. Using systems and structure biology tools to dissect cellular phenotypes. *J Am Med Inform Assoc* 2012;**19**:171–5.
8. **Musen MA,** Noy NF, Shah NH, *et al*. The National Center for Biomedical Ontology. *J Am Med Inform Assoc* 2012;**19**:190–5.
9. **Kohane IS,** Churchill SE, Murphy SN. A translational engine at the national scale: informatics for integrating biology and the bedside. *J Am Med Inform Assoc* 2012;**19**:181–5.
10. **Liu M,** Wu Y, Chen Y, *et al*. Large-scale prediction of adverse drug reactions by integrating chemical, biological, and phenotypic properties of drugs. *J Am Med Inform Assoc* 2012;**19**:e28–35.
11. **Bhavnani SK,** Bellala G, Victor S, *et al*. The Role of Complementary Bipartite Visual Analytical Representations in the Analysis of SNPs: A Case Study in Ancestral Informative Markers. *J Am Med Inform Assoc* 2012;**19**:e5–12.
12. **Russu A,** Malovini A, Puca AA, *et al*. Stochastic model search with binary outcomes for Genome-Wide Association Studies. *J Am Med Inform Assoc* 2012;**19**:e13–20.
13. **Morgan AA,** Chen R, Butte AJ, *et al*. Clinical utility of sequence-based genotype compared with that derivable from genotyping arrays. *J Am Med Inform Assoc* 2012;**19**:e21–7.
14. **Kho AN,** Hayes MG, Rasmussen-Torvik L, *et al*. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *J Am Med Inform Assoc* 2012;**19**:212–18.
15. **Worthey EA,** Mayer AN, Syverson GD, *et al*. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet Med* 2011;**13**:255–62.
16. **Mayer AN,** Dimmock DP, Arca MJ, *et al*. A timely arrival for genomic medicine. *Genet Med* 2011;**13**:195–6.
17. **Ashley EA,** Butte AJ, Wheeler MT, *et al*. Clinical assessment incorporating a personal genome. *Lancet* 2010;**375**:1525–35.
18. **Samani NJ,** Tomaszewski M, Schunkert H. The personal genome—the future of personalised medicine? *Lancet* 2010;**375**:1497–8.
19. **Harpaz R,** *et al*. Novel data mining methodologies for adverse drug event discovery and analysis. *Nature Clinical Pharmacology and Therapeutics*. In press (accepted 5 Mar 2012).
20. **Bate A,** Evans SJW. Quantitative signal detection using spontaneous ADR reporting. *Pharmacoepidemiol Drug Saf* 2009;**18**:427–36.
21. **Szarfman A,** Machado SG, O'Neill RT. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database. *Drug Saf* 2002;**25**:381–92.
22. **Weiss-Smith S,** Deshpande G, Chung S, *et al*. The FDA drug safety surveillance program: adverse event reporting trends. *Arch Intern Med* 2011;**171**:591–3.
23. **Norén N,** Sundberg R, Bate A, *et al*. A statistical methodology for drug-drug interaction surveillance. *Stat Med* 2008;**27**:3057–70.
24. **Tatonetti NP,** Denny JC, Murphy SN, *et al*. Detecting drug interactions from adverse-event reports: interaction between paroxetine and pravastatin increases blood glucose levels. *Clin Pharmacol Ther* 2011;**90**:133–42.
25. **Tatonetti NP,** Ye PP, Daneshjou R, *et al*. Data-driven prediction of drug effects and interactions. *Sci Transl Med* 2012;**4**:125–31.
26. **Cami A,** Arnold A, Manzi S, *et al*. Predicting adverse drug events using pharmacological network models. *Sci Transl Med* 2011;**3**:114–27.
27. **Pouliot Y,** Chiang AP, Butte AJ. Predicting adverse drug reactions using publicly available PubChem BioAssay data. *Clin Pharmacol Ther* 2011;**90**:90–9.
28. **Vilar S,** Harpaz R, Chase HS, *et al*. Facilitating adverse drug event detection in pharmacovigilance databases using molecular structure similarity: application to rhabdomyolysis. *J Am Med Inform Assoc* 2011;**18**(Suppl 1):i73–80.
29. **Liu Y,** LePendu P, Iyer S, *et al*. Using temporal patterns in medical records to discern adverse drug events from indications. *AMIA Summit on Clinical Research Informatics*. San Francisco: AMIA, 2012.
30. **LePendu P,** Iyer SV, Fairon C, *et al*. Annotation analysis for testing drug safety signals. *Journal of Biomedical Semantics* 2012;**3**(Suppl 1):S5.
31. **Brownstein JS,** Sordo M, Kohane IS, *et al*. The tell-tale heart: population-based surveillance reveals an association of rofecoxib and celecoxib with myocardial infarction. *PLoS One* 2007;**2**:e840.
32. **Harpaz R,** Chase H, Friedman C. Mining multi-item drug adverse effect associations in spontaneous reporting systems. *BMC Bioinformatics* 2010;**11** (Suppl 9):S7.
33. **Dore D,** Seeger J, Arnold Chan K. Use of a claims-based active drug safety surveillance system to assess the risk of acute pancreatitis with exenatide or sitagliptin compared to metformin or glyburide. *Curr Med Res Opin* 2009;**25**:1019–27.
34. **Nadkarni P.** Drug safety surveillance using de-identified EMR and claims data: issues and challenges. *J Am Med Inform Assoc* 2010;**17**:671–4.
35. **Brown JS,** Kulldorff M, Chan KA, *et al*. Early detection of adverse drug events within population-based health networks: application of sequential testing methods. *Pharmacoepidemiol Drug Saf* 2007;**16**:1275–84.
36. **Shetty KD,** Dalal S. Using information mining of the medical literature to improve drug safety. *J Am Med Inform Assoc* 2011;**18**:668–74.
37. **Chee BW,** Berlin R, Schatz B. Predicting adverse drug events from personal health messages. *AMIA Annu Symp Proc* 2011;**2011**:217–26.
38. **Sobek M,** Cleveland L, Flood S, *et al*. Big data: large-scale historical infrastructure from the Minnesota Population Center. *Hist Methods* 2011;**44**:61–8.
39. **Fox B.** Using big data for big impact. How predictive modeling can affect patient outcomes. *Health Manag Technol* 2012;**33**:32.
40. **Chen R,** Mias GI, Li-Pook-Than J, *et al*. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* 2012;**148**: 1293–307.
41. **Frankovich J,** Longhurst CA, Sutherland SM. Evidence-based medicine in the EMR era. *N Engl J Med* 2011;**365**:1758–9.
42. **Halevy A,** Norvig P, Pereira F. The unreasonable effectiveness of data. *IEEE Intelligent Systems* 2009;**24**:8–12.
43. **Hays J,** Efros AA. Scene completion using millions of photographs. *Commun ACM* 2008;**51**:87–94.
44. **Bringardner J.** Winning the lawsuit: data miners dig for dirt. *Wired Magazine* 2008:16–07.
45. **Michel J-B,** Shen YK, Aiden AP, *et al*. Quantitative analysis of culture using millions of digitized books. *Science* 2011;**331**:176–82.
46. **CBO.** *The Long Term Outlook for Health Care Spending*. 2007. http://www.cbo.gov/sites/default/files/cbofiles/ftpdocs/87xx/doc8758/11-13-lt-health.pdf
47. **Lorenzi NM.** *AMIA's Realigned Strategic Plan*. 2011. http://www.amia.org/issues/amias-realigned-strategic-plan (accessed 13 Mar 2012).
48. **Friedman CP,** Wong AK, Blumenthal D. Achieving a nationwide learning health system. *Sci Transl Med* 2010;**2**:57cm29.
49. **LePendu P,** Musen MA, Shah NH. The age of data-driven medicine: mining the electronic health record. *International Conference on Biomedical Ontologies*. Buffalo, NY: CEUR, 2011:435.