Instrumenting the health care enterprise for discovery research in the genomic era

Shawn Murphy, Susanne Churchill, Lynn Bry, et al.

*Genome Res.* 2009 19: 1675-1681 originally published online July 14, 2009
Access the most recent version at doi:10.1101/gr.094615.109

**References**
This article cites 24 articles, 5 of which can be accessed free at:
http://genome.cshlp.org/content/19/9/1675.full.html#ref-list-1

Article cited in:
http://genome.cshlp.org/content/19/9/1675.full.html#related-urls

**Open Access**
Freely available online through the *Genome Research* Open Access option.

**Creative Commons License**
This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported License), as described at http://creativecommons.org/licenses/by-nc/3.0/.

**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 article or click here

To subscribe to *Genome Research* go to:
http://genome.cshlp.org/subscriptions

© 2009, Published by Cold Spring Harbor Laboratory Press
Instrumenting the health care enterprise for discovery research in the genomic era

Shawn Murphy,1,2 Susanne Churchill,2 Lynn Bry,3 Henry Chueh,4 Scott Weiss,5 Ross Lazarus,5 Qing Zeng,6 Anil Dubey,1 Vivian Gainer,1 Michael Mendis,1 John Glaser,2,7,8 and Isaac Kohane2,8,9,10,11

1Informatics, Partners Healthcare Systems, Boston, Massachusetts 02115, USA; 2i2b2 National Center for Biomedical Computing, Boston, Massachusetts 02115, USA; 3Department of Pathology, Brigham and Women’s Hospital, Boston, Massachusetts 02115, USA; 4Massachusetts General Hospital Laboratory for Computer Science, Boston, Massachusetts 02114, USA; 5Channing Laboratory, Brigham and Women’s Hospital, Boston, Massachusetts 02115, USA; 6Decision Systems Group, Brigham and Women’s Hospital, Boston, Massachusetts 02115, USA; 7Information Systems, Partners Healthcare Systems, Boston, Massachusetts 02115, USA; 8Department of Medicine, Brigham and Women’s Hospital, Boston, Massachusetts 02115, USA; 9Children’s Hospital Informatics Program at the Harvard–Massachusetts Institute of Technology Division of Health Sciences and Technology, Boston, Massachusetts 02115, USA; 10Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts 02115, USA

Tens of thousands of subjects may be required to obtain reliable evidence relating disease characteristics to the weak effects typically reported from common genetic variants. The costs of assembling, phenotyping, and studying these large populations are substantial, recently estimated at three billion dollars for 500,000 individuals. They are also decade-long efforts. We hypothesized that automation and analytic tools can repurpose the informational byproducts of routine clinical care, bringing sample acquisition and phenotyping to the same high-throughput pace and commodity price-point as is currently true of genome-wide genotyping. Described here is a demonstration of the capability to acquire samples and data from densely phenotyped and genotyped individuals in the tens of thousands for common diseases (e.g., in a 1-yr period: N = 15,798 for rheumatoid arthritis; N = 42,238 for asthma; N = 34,535 for major depressive disorder) in one academic health center at an order of magnitude lower cost. Even for rare diseases caused by rare, highly penetrant mutations such as Huntington disease (N = 102) and autism (N = 756), these capabilities are also of interest.

A common thread in the recent flurry of studies relating characteristics of complex diseases to the generally weak effects of individual genetic variants is that very large numbers of subjects are needed to obtain reproducible results—closer to 200,000 individuals (Manolio et al. 2006) than the few thousand typical of recent publications. The costs of assembling, phenotyping, and studying these huge populations are estimated at three billion dollars for 500,000 individuals (Spivey 2006). Reciprocally, studying rare diseases often requires searching through very large populations, and sufficient sample sizes are hard to achieve. Coincidentally, the United States spends over two trillion dollars in healthcare per year (Caflin et al. 2008), and of those costs, the total investment in information technology (IT) is at least seven billion dollars per year (Girosi et al. 2005). The stimulus package recently enacted by the U.S. Congress includes a very significant increase in spending on electronic health records, prompting interest in the secondary use of the data gathered in such records. Yet there is widespread, often justified skepticism about our ability to use routinely collected electronic health records (EHRs) for research-quality phenotype data, given the well-known biases and coarse-grained nature of billing/claims diagnoses and procedures (Safran 1991; Jollis et al. 1993). By the same measure, the consistency of phenotypic definitions in large genome-wide association studies (GWAS), especially when they consist of the aggregation of several existing studies, and the consequent effect upon these study results, has been questioned (Ioannidis 2007; Wojcynski and Tiwari 2008; Buyske et al. 2009).

To meet these challenges, we have undertaken a series of institutional experiments that collectively demonstrate that automated systems for mining of EHRs are essential for the feasibility and affordability of large-scale population studies such as GWAS. We do so by using a free and open-source system, i2b2 (Informatics for Integrating Biology and the Bedside; http://www.i2b2.org) to conduct a proof-of-principle exercise to show that this system (1) accurately identifies potential cases and controls by mining the EHR using natural language processing (NLP), and it does this (2) much faster and (3) much more cheaply than traditional methods.

Methods

A central goal of i2b2 is to test our methodologies with “Driving Biology Projects” (DBP) led by investigators interested in specific disease areas (e.g., pharmacogenomics of asthma, risk alleles for rheumatoid arthritis [RA], and variants associated with resistance to the antidepressant effects of selective serotonin reuptake inhibitors). We outline here the general approach to a DBP and then illustrate it with specifics from two DBPs.

First, the investigators select a set of terms that are used routinely in clinical practice to diagnose or stage a condition (e.g., asthma), preferably including findings that are part of the “standard” classification criteria for that disease. These terms are augmented with those medications that are specific to the diseases of interest. Also considered are those diseases or conditions that are frequent mimickers of the disease of interest to define terms that should be excluded.

E-mail isaac_kohane@hms.harvard.edu; fax (206) 333-1182.
Once the term list has been developed, it is submitted to the NLP utility of i2b2. This utility, called HITEx (Zeng et al. 2006), is built upon the popular and open-source GATE (Cunningham et al. 2002) framework from the University of Sheffield. HITEx then operates over the millions of clinical narratives (e.g., discharge summaries, clinic notes, preoperative notes, pathology and radiology reports) in the EHR and generates a set of codified concepts drawn from the Unified Medical Language System (Lindberg and Humphreys 1992). Each of these concepts is entered into the same database that contains all the pre-existing institutional EHR clinical data (e.g., laboratory studies, billing codes) and labeled as derived data. Regardless of their origin (i.e., primary data or derived data), the entire database can then be searched to find sets of patients that meet specified criteria such as comorbidities (e.g., bronchitis), exposures (e.g., smoking), medications taken, or laboratory results (e.g., positive anticlyclic citrullinated peptide antibody assays).

DBP clinical experts are then recruited to review the results of queries using the concepts individually (whether NLP-defined or codified originally in the EHR) and combined for accuracy. This is done by reading the full clinical narrative text corresponding to a random subsample of patients selected by these queries to establish the "gold-standard" phenotype for those patients. Then, regression methods are applied to train prediction models that relate the variables to the phenotype of interest. When the number of available variables is not small, regularized regression procedures with an adaptive lasso (Tibshirani 1996) penalty are employed to identify important features and train the final model for prediction with the selected variables. Based on a separate validation data set, the prediction performance using measures including the receiver operating characteristic (ROC) curve, the positive and negative predictive values are assessed.

The sample size of the training data is determined adaptively. We first randomly select an initial set of records for review to train the model. With the same set of data, we obtain initial confidence interval estimates of the predictive accuracy using the cross-validation and bootstrap method. Subsequently, we determine the required sample size for both the training and the validation data sets based on the desired width of the confidence intervals. Typically, the training and validation data sets require review of the records of <500 patients by the clinician experts.

Once the selection methods are fine-tuned, the selected group of patients is retrieved and, per our institutional review board (IRB) protocol, that database is "frozen" as a "datamart" for that DBP. From that datamart, a set of unique, anonymous identifiers is generated. As illustrated in Figure 1, we then ran a trial using Crimson, a new resource developed by the Department of Pathology at Brigham and Women’s Hospital, which offers IRB-compliant access to discarded blood samples for genotyping. Patient identifiers extracted using i2b2 in silico phenotyping are forwarded to the Crimson application. The Crimson application queries recently accessioned materials from clinical patient visits against the i2b2-forwarded identifiers. Instead of being discarded, matching samples are accessioned into Crimson, with the sample assigned to the requesting study’s IRB protocol, and the patient identifier converted to a unique anonymized i2b2 code. Crimson generates an anonymous sample identifier so that no original identifiers (laboratory accession number, medical record number, etc.) remain associated with the sample, which can be released for DNA extraction and further analysis, with a rich set of previously extracted and deidentified phenotypes from the medical record system.

The anonymity described here is highly circumscribed and critically dependent on institutional review. All Health Insurance Portability and Accountability Act (HIPAA)-described identifiers are removed, and all codes linking the record to the patient identity are deleted. Also, any systematic attempt of re-identification is strictly prohibited and is a violation of IRB protocol resulting in severe penalties to the investigators who also are employees of the healthcare system.

The first DBP to successfully employ the process described above was the asthma DBP. The project focused on acute asthma exacerbations requiring hospitalization, because these are a major cause of health care costs for asthma and these events are readily identified through the pre-existing research patient data repository. The asthma DBP had previously defined clinical and genetic predictors of asthma hospitalizations based on a GWAS conducted in an independent cohort. The study goal was to select the cases (high utilizers) and controls (low utilizers) and confirm the previously identified genetic predictors of hospitalizations.
The results described here pertain to the 2.6 million patients seen at the two major hospitals within the Partners Healthcare System.
(the Brigham and Womens’ Hospital and the Massachusetts General Hospital), of which 821,925 are seen per year, generating over 3,300,000 tubes of blood per year.

**Accrual rates (forecast and actual)**

The i2b2 toolkit provides a mechanism for both patient accrual and forecasting the rate of accrual for any cohort of interest. For example, in the instance of an asthma study, we predicted an accrual of 3174 patients fitting our case and control definitions of utilization in the one hospital (based on how many with the same phenotypic definition had returned for care the prior year). Figure 2 shows the actual accrual sample from asthma subjects stratified by high healthcare services utilizers and low healthcare services utilizers (as defined above) and by race (African American and Caucasian American). Figure 3 shows the projected accrual rate in other example diseases or syndromes, including all individual with asthma, not just those meeting our particular study criteria. Even in a midsized academic healthcare center, thousands of phenotyped samples can be acquired for common diseases at a rate of over 300 per week. Even when the goal is identification of rare diseases, where a few hundred patients would enable an important study, this system allows hundreds of thousands of patients to be efficiently phenotyped so that these rare cases can be identified and their samples obtained (as in Huntington disease in Fig. 3). It can also be used to identify rare events such as Steven Johnson syndrome (5284 cases returned to the health system this year of those identified in prior years) to allow genomic study of such events.

**Phenotyping**

In the asthma DBP, HITEx was used to extract principal diagnosis, comorbidity, and smoking status from discharge summaries and outpatient visit notes as described above. Unlike some NLP packages, HITEx will report for each possible disease not only whether it is present or absent but also if there are “insufficient data” to reach a sound conclusion. To compare HITEx results to the human ratings, we treated the “insufficient data” label in three ways: excluding cases with that label, treating them as “present,” and treating them as “absent.”

Accuracy was evaluated for the asthma DBP in random samples by experienced pulmonologists reviewing the full medical record. Compared with the experts, the accuracy of the i2b2 NLP program HITEx (Zeng et al. 2006) for principal diagnosis extraction was 73%–82% and for comorbidity was 78%–87%, depending on how the expert label “insufficient data” was treated. HITEx accuracy was 1%–4% higher than the expert analysis using the ICD-9 diagnosis code in every category. This relative measure obviously only makes sense where there is an ICD-9 code that actually corresponds to a concept obtained by NLP. The accuracy of HITEx smoking status extraction was 90%. However, this performance was a result of an iterative process between domain experts (e.g., pulmonologists) and the NLP experts, without which, using current technology, the outcome would be much less satisfactory. In subsequent DBPs we have been able to consistently attain accuracies of over 92% (for RA and major depressive disorder resistant to selective serotonin reuptake inhibitors).

Figure 4 illustrates the challenge by providing a glimpse of just how heterogeneous the human-driven characterizations are for merely one attribute: smoking history. Nonetheless, once the HITEx package is tuned, running it against millions of patient reports is just a matter of days with the accuracies reported here. In contrast, medical chart review by even a non-expert (e.g., medical student) takes 15 min (and easily several times that with more complex charts) at a cost of $20 per record reviewed.

The RA investigators systematically identified the features of interest (HITEx-derived and also previously codified) using a logistic regression approach with the adaptive lasso penalty. They identified seven predictors of RA using their gold-standard set of RA patients and non-RA patients: disease codes for RA and three diseases that mimic RA, NLP-derived medication annotations, and NLP-derived seropositivity. This RA selection algorithm was used to select patients from the entire datamart. A total of 4618 subjects were selected, and 92% of patients had definite RA and 98% had either probable or definite RA. Of note, over 40% of the ostensible cases of RA in the datamart were due to quirks in the codification/billing process (e.g., radiologists codifying a “rule-out” RA with the RA ICD-9 billing code). When the NLP-derived medication
records were compared with those in the codified entries; 98% of patients who had an electronic prescription also had a HITEx annotation for the medication of interest. Conversely, HITEx identified twice as many RA medications as reported by the electronic prescription data.

Costs

Figure 5 illustrates a projection of the costs of a GWAS for study populations ranging in size from one thousand to one million. The projections cover a wide range of cost assumptions (see Methods). This result concurs with the published estimates for one million patients, which are well into the nine-figure range (Spivey 2006). It also illustrates how judicious use of state-of-the-art technologies for phenotyping and sample acquisition can reduce the cost of these studies by half an order of magnitude (from $1.2 billion to $520 million). The implementation of $100/sample genome-wide variant assays brings that same cohort cost down another half order of magnitude to $150 million. These projections might be further modified if there were economies of scale through automation to reduce the per sample costs, an assumption not included in these conservative models.

These estimates assume a very significant pre-existing infrastructure for the purposes of providing high-quality care. This includes an electronic health record (Committee on Quality of Health Care in America, Institute of Medicine 2001) and data warehouse, a high-volume clinical laboratory information system, and competent, engaged information systems staff. All these investments are typically made for reasons other than supporting discovery research so they are not included in i2b2 cost estimates. The generic “star schema” (Kimball and Ross 2002) of the i2b2 datamart supports a wide variety of clinical and genomic data types. This in turn has allowed IT staff from across the more than 36 implementation sites (of which five are outside the United States; see https://www.i2b2.org/work/aug.html) to import data from their EHRs, including locally developed systems as well as commercial offerings from Cerner Corporation, Meditech Information Technology, NextGen Health Information Systems, and Epic Systems Corporation.

Conclusion

The approach described is not without limitations. Despite a multiplicity of blue-ribbon panels and reports (Committee on Quality of Health Care in America, Institute of Medicine 2001) on the improvement in the quality of care that results, less than 20% of healthcare enterprises currently have suitable information infrastructure (Poon et al. 2006), although this may grow significantly with the recent passage of the Health Information Technology for Economic and Clinical Health (HITECH) Act (Senate and House of Representatives of the United States of America in Congress 2009). Even if
These estimates are conservative, as none of the models considered provide for any improved efficiencies of scale.

Costs of instrumenting the healthcare enterprise. Growth in costs of study as a function of number of subjects in a study is projected for different assumptions of the cost of sample acquisition, phenotyping, and genotyping. Eight lines are drawn corresponding to eight combinations of these three costs. The main diagram shows the projection for up to 20,000 subjects and the inset for up to one million. The costs for sample acquisition using i2b2 sample acquisition are $20 (LS) or $50 per sample for a larger population (HS) vs. the current cost (CS) of $650. The current costs for reviewing one record to phenotype a patient (CP1) or, more typically, five records reviewed per study patient identified (CP2) are estimated at $20/sample and $100/sample, respectively. High-throughput phenotyping through NLP (IP) is conservatively estimated at $50,000 per study. Current cost of genome-scale genotyping (CG) vs. lower cost genotyping (LG) within three years is estimated at $500 vs. $100, respectively. There is a range of about one-half order of magnitude cost reduction from having the phenotyping and sample acquisition done using i2b2 and another order of magnitude using genotyping costs projected for no more than three years from now. This is a difference, for a million-subject study, that covers a range from $1.2 billion to $150 million. These estimates are conservative, as none of the models considered provide for any improved efficiencies of scale.

Concerns about the risks to patient privacy or the appearance of risk are barriertoe widespread use of electronic health care data for research. Regulatory protection of patient privacy should, in principle, not obstruct or unduly retard the conduct of clinical research, although in practice the principle is often obscured (O’Herrin et al. 2004). Clearly, cavalier handling of such data sets can lead to real risks (Russell and Theodore 2005; United States Congress Senate Committee on Veterans’ Affairs and United States Congress Senate Committee on Homeland Security and Governmental Affairs 2007) even while the practice of medicine itself remains highly disclosing of patient information (Clayton et al. 1997; Sweeney 1998). Moreover, most genome-wide data is highly disclosing (Homer et al. 2008) and the public release of such data is fraught with risks to privacy. This is a challenge that any study involving GWAS, whether or not it uses i2b2, must address.

With the increased impetus toward the implementation of electronic health records and the intense interest in evaluating genomic tools and applications, there is a means to render such research more timely and cost-effective. With the increased impetus toward the implementation of electronic health records and the intense interest in evaluating genome-scale signatures in large populations, the time is ripe for wider adoption of such methods.

References


Received April 5, 2009; accepted in revised form July 13, 2009.