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 Kohane[2,8,9,10,11]

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

[11]**Corresponding author.**
E-mail isaac_kohane@hms.harvard.edu; fax (206) 333-1182.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.094615.109. Freely available online through the *Genome Research* Open Access option.

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 anticyclic 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.

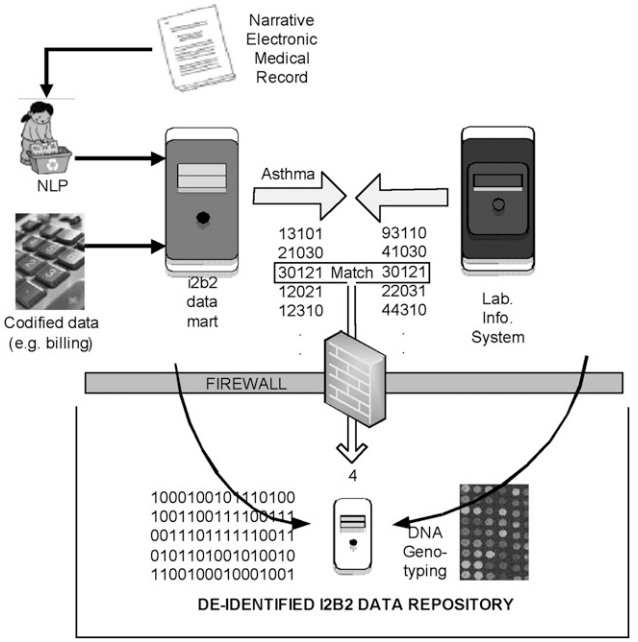

Figure 1. Matching anonymously identified populations to anonymous samples. An i2b2 datamart is generated from codified data (e.g., billing codes, laboratory test values) and concepts codified by running the narrative text in electronic medical records through a NLP tool, the HITEx package described in the Methods section. Patients included within the i2b2 datamart meeting study criteria are selected and their corresponding set of identifiers are generated. Those identifiers are forwarded to the Crimson application, which scans recent transactions forwarded from one or more local clinical laboratory or pathology information systems to identify newly accessioned materials matching the cohort identifiers and desired sample types. Upon completion of diagnostic testing (1–3 d after collection in most cases), Crimson manages reaccession of the sample to a study's IRB protocol and assigns the i2b2-forwarded subject ID to a uniquely generated sample ID. These actions remove all identifiers (accession no., medical record no., etc.) from the original sample. The sample may then be released to the investigator where it can be measured (for genome-wide genotyping in this instance), and these measurement data are merged with the phenotypic data set in the i2b2 datamart. Because of the electronic and regulatory firewalls, only research personnel approved by the IRB can view the limited data set (in the HIPAA sense), and they cannot view the identified clinical data visible to those who access the laboratory information 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.

Table 1. Gold-standard task for expert reviewer in the asthma DBP

1. Principal diagnosis includes asthma: yes/no/insufficient data
2. Principal diagnosis includes COPD: yes/no/insufficient data
3. Comorbidities include asthma: yes/no/insufficient data
4. Comorbidities include COPD: yes/no/insufficient data
5. Smoking status: current smoker, past smoker, nonsmoker, patient denies smoking, insufficient data

These are the annotations that each reviewer had to provide for each record reviewed. The annotation types were selected by the pulmonologists by virtue of their relevance to sample selection for the GWAS and subsequent analyses.

The second DBP to complete the use of i2b2 phenotyping was the RA DBP that has as its goal a GWAS study of RA cases versus controls to confirm prior findings and find new risk alleles.

Phenotyping

Phenotypic characterization of the 97,639 patients in the asthma cohort used the NLP package (HITEx) as described above to stratify patients by smoking history, healthcare utilization, severity, and medications. Other phenotypes captured included detailed pulmonary function test results (extracted from textual reports), comorbidities, and family history. The gold-standard annotation was established by expert review of a random sample of clinical reports to answer the questions in Table 1 for each report. HITEx then was run to evaluate the same five questions for each report.

In the asthma DBP, subjects with a doctor's diagnosis of asthma, aged >15 yr and <45 yr, who were non- or ex-smokers, and who had no hospitalizations (low utilizer group) or greater than two hospitalizations (high utilizer group) within a 36-mo interval were selected from the 97,639 asthmatics identified through the initial "high-throughput" phenotyping.

The RA i2b2 datamart includes patients seen between October 1993 and June 2008 with any ICD-9 diagnostic code for RA or a related condition. Using an independent prospective annotation of 1025 RA patients recruited for a previous study at Partners Healthcare Systems, we found that these criteria were highly sensitive for the diagnosis of RA (99% of RA patients were included in their RA datamart). Two clinical rheumatologists reviewed 500 randomly selected charts and identified a gold-standard set of RA patients and non-RA patients (102 definite RA cases and 398 non-RA patients).

Costs for large-scale i2b2 association studies

We initially opted to collaborate with the physicians in the largest outpatient clinics to recruit and consent their patients as they appeared for routine care. Unfortunately, this yielded fewer than 10 patients per week, with costs of about $650 per patient sample. This is a cost comparable to the typical reported range of $500–$1200, without phlebotomy costs, for noncommercial population studies (Gismondi et al. 2005; Ota et al. 2006; Karlawish et al. 2008) but was too expensive and slow for our purposes. These high costs and slow recruitment rate led to our aforementioned development of the link between i2b2 and the Crimson system.

If we presume a very significant prior and ongoing investment in an IT infrastructure (for quality clinical care) and discount the analytic steps that are shared in all studies, regardless of how the study materials are accumulated, the incremental costs of each new study can be categorized within three categories: the costs of phenotyping, the costs of sample acquisition, and the costs of genome-scale measurements (summarized in Table 2).

Current sample acquisition as practiced in most studies costs upwards of $650 per patient. i2b2 sample acquisition currently is under $20 per sample including DNA extraction costs. For larger populations, additional infrastructure for storage and retrieval might push this cost as high as $50 per sample. Current phenotyping costs through manual chart review are a function of how many records will have to be reviewed to obtain a single phenotyped patient. Current phenotyping costs conservatively average $20 (Allison et al. 2000; Flynn et al. 2002), whereas the costs at the higher estimate for current phenotyping are conservatively estimated at $100 per patient identified, that is, five charts reviewed for every patient included in the study. Both the lower and higher estimates of current phenotyping costs are assumed to scale linearly with the numbers of patients sought.

i2b2 phenotyping requires a substantial initial investment in defining the phenotypes of interest, "tuning" the NLP methods iteratively. This multidisciplinary team effort currently entails an additional investment, mostly in analytic personnel costs, of $20,000 to $50,000, but this range is largely independent of the sample size sought and can be run multiple times across the years at nominal incremental costs. We use the higher cost estimate of i2b2 phenotyping in the calculations below.

If we take the current practice of measurement of common variants as the standard for genome-wide studies, then the cost of genomic measurements, including labor and materials, is no more than $500 per patient (2008). Based on past performance and current predictions, genome-wide genotyping costs are likely to drop to less than $100 within the next three years.

Results

The results described here pertain to the 2.6 million patients seen at the two major hospitals within the Partners Healthcare System

Table 2. Dollar and time costs

	Cost ($)	Time cost
One chart review per patient (CP1)	20	15 min/subject
Five chart reviews for one subject (CP2)	100	45 min/subject
High-throughput phenotyping (iP)	50,000	1 mo total (conservative high estimate)
Sample acquisition through primary care provider (CP)	650	3–5 subjects/wk[a]
High-throughput sample acquisition, lower cost (LP)	20	50–200 subjects/wk[b]
High-throughput sample acquisition, higher cost (HP)	50	50–200 subjects/wk[b]
Current genome-wide SNP scan	500	20 samples/d
Future genome-wide SNP scan	100	100 samples/d

The cost of sample acquisition, phenotyping and genotyping in dollars used for the models illustrated in Fig. 5. The three costs for sample acquisition costs are the low and high costs using i2b2 per sample versus the current cost (denoted LS, HS, CS). The current cost for reviewing one record to phenotype a patient (CP1) or, more typically, five records reviewed per study patient identified are denoted CP1 and CP2 and i2b2 phenotyping as iP. Current cost genome-scale genotyping versus lower cost genotyping within three years are denoted CG and LG.
[a]Data from Asthma Driving Biology Project (DBP).
[b]See Figure 2.

(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.

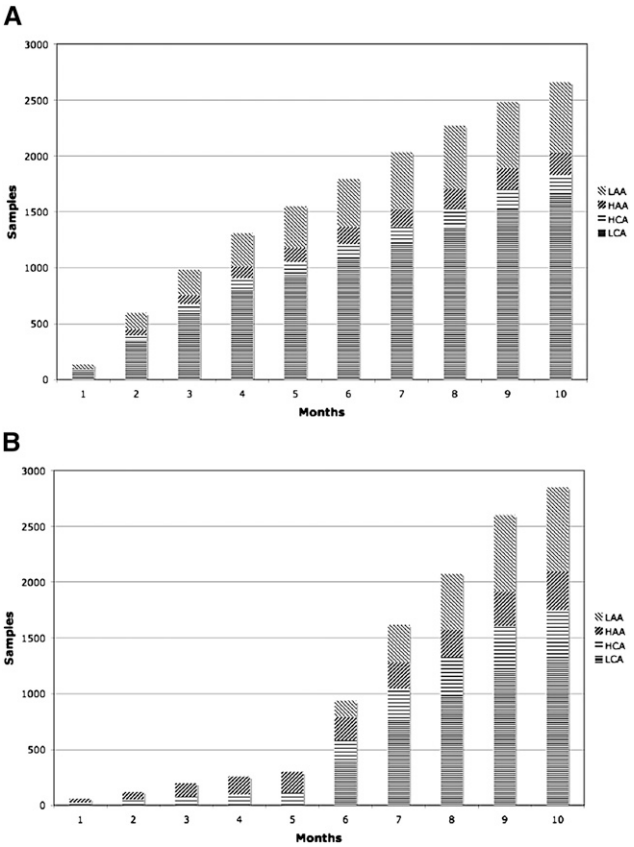

Figure 2. Cumulative accrual of phenotyped DNA samples for the asthma DBP. Unlike the membership of the overall asthma datamart (N = 131,230), the pool from which patients were drawn was first restricted to those seen at the Brigham and Women's Hospital where the Crimson system was first deployed, as the second Crimson site (Massachusetts General Hospital) came online only in late 2008. Additional restrictions included age (<45 yr, >15 yr) and smoking status (nonsmoker or ex-smoker). Projected (*A*) and actual (*B*) accrual rates for four groups: LCA (Low-utilizer Caucasian American); LAA (Low-utilizer African American); HCA (High-utilizer Caucasian American); HAA (High-utilizer African American). Utilization here is defined by the absence of hospital admissions (low utilizer) in contrast to at least two admissions (high utilizer). As shown above, the recruitment of low utilitzers (LCA and LAA) started later than the recruitment of the high utilizers. Nonetheless, the projected recruitment rates and the actual recruitment rates are very similar.

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 as having a high probability of RA (at 97% specificity). Of those, a random sample of 400 charts from these 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

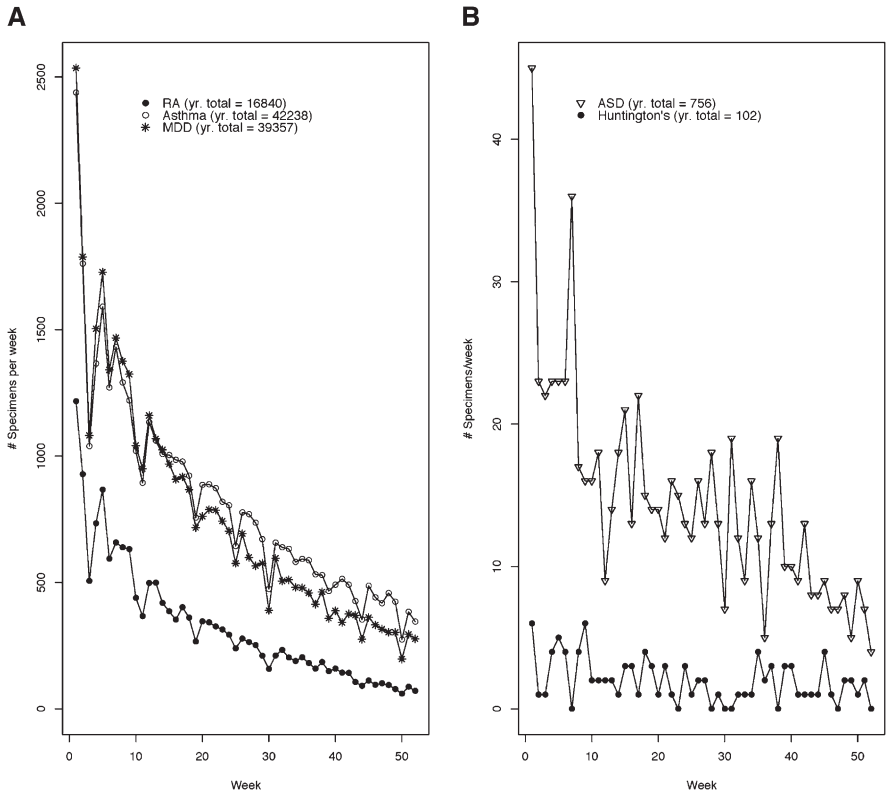

Figure 3. Projected accrual rates. Estimates are based on the number of patients previously seen at least once during the 36 mo before June 30, 2006 for whom at least one patient visit during which chemistry or hematology samples were obtained was then recorded in the following 12 mo. Each patient was only counted once, even if they had more than one visit in the 12-mo period. Also, unlike Figure 2, accrual rates per week rather than cumulative accrual are shown. There are some common features in the accrual trajectories for most of the diseases because of the shared exposure to the effects of holidays and seasonality on hospital visits. (*A*) Accrual for common diseases: (MDD) major depressive disorder; (RA) rheumatoid arthritis (all individuals and not just those who met driving biological projects criteria); asthma (also all individuals). (*B*) Accrual for less prevalent diagnoses: Huntington disease and autism spectrum disorder (ASD) (including Asperger syndrome).

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

Medical Record Snippet	*Smoking History*
SOCIAL HISTORY: The patient is married with four grown daughters, uses tobacco, has wine with dinner.	Positive
SOCIAL HISTORY: The patient is a nonsmoker. No alcohol.	Negative
SOCIAL HISTORY: Negative for tobacco, alcohol, and IV drug abuse.	Negative
BRIEF RESUME OF HOSPITAL COURSE: 63 yo woman with COPD, 50 pack-yr tobacco (quit 3 wks ago), spinal stenosis, ...	Positive
SOCIAL HISTORY: The patient lives in rehab, married. Unclear smoking history from the admission note…	Insufficient data
HOSPITAL COURSE: ... It was recommended that she receive …We also added Lactinax, oral form of Lactobacillus acidophilus to attempt a repopulation of her gut.	Insufficient data
SH: widow,lives alone,2 children,no tob/alcohol.	Insufficient data

Figure 4. Example smoking annotations in electronic medical records. The boxes around selected words highlight those the HITEx system picked up as informative regarding smoking status. The second column provides the system's classification of the smoking status. This illustrates the challenges for which additional tuning was required. For example, the "tobac" in *Lactobacillus* is no less obvious to HITEx, initially, than the "tob" in "tob/alcohol."

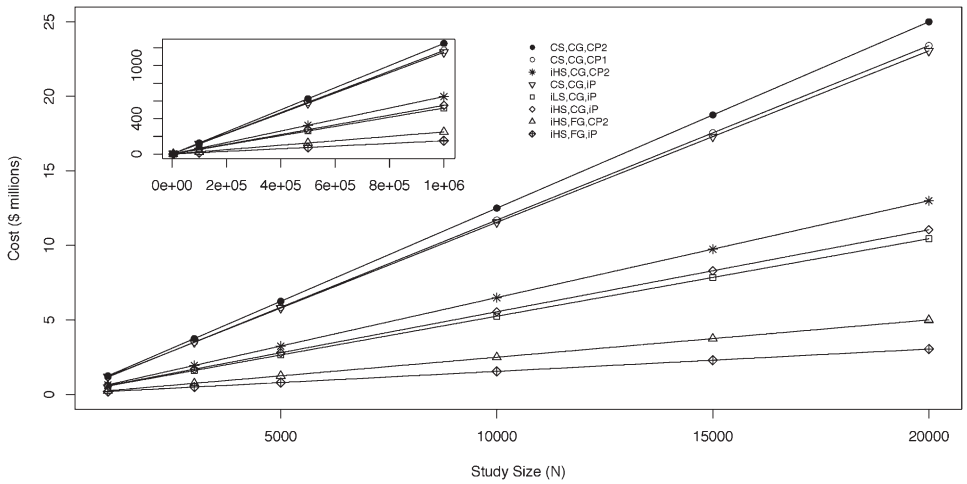

Figure 5. Costs of instrumenting the healthcare enterprise. Growth in costs of study as a function of number of subjects in a study is projected for different assumptions of the cost of sample acquisition, phenotyping, and genotyping. Eight lines are drawn corresponding to eight combinations of these 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.

phenotype information continues to accrue, many important measures of health and environment will likely remain absent from the institutional/provider-driven health record, although mechanisms such as personally controlled health records (Kohane et al. 2007) may eventually help fill this gap. Patients who have the opportunity to correct or enhance existing medical records (Porter et al. 2000) often have the most to gain from such corrections. With regard to demographic representation, accrual results (Fig. 2) show that minorities are over-represented compared with local demographics, confirming that patients of an academic medical center may differ from the general population in important ways.

Concerns about the risks to patient privacy or the appearance of risk are barriers to 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 regard to the use of discarded anonymous specimens for the sample acquisition, we note that the machinery described here can be used to prospectively cast a broad net for consented samples among patient groups and then use NLP to identify suitable samples. This corresponds to the operation of Vanderbilt University's BioVU system (Roden et al. 2008), where all patients are offered an "opt-out" check box on each of the standard forms they sign to obtain healthcare. In its current operation, unlike BioVU, i2b2's datamarts and biorepositories are created "on demand" for investigators. To date, this has scaled well when mining healthcare systems with several million patients for populations of interest numbering in the thousands or tens of thousands.

Finally, i2b2 is best understood as one of the consequences of a logical progression of over four decades of clinical research (Warner 1966; Safran et al. 1989) using electronic health records as 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

Allison JJ, Wall TC, Spettell CM, Calhoun J, Fargason CA Jr, Kobylinski RW, Farmer R, Kiefe C. 2000. The art and science of chart review. *Jt Comm J Qual Improv* **26:** 115–136.

Buyske S, Yang G, Matise T, Gordon D. 2009. When a case is not a case: Effects of phenotype misclassification on power and sample size requirements for the transmission disequilibrium test with affected child trios. *Hum Hered* **67:** 287–292.

Catlin A, Cowan C, Hartman M, Heffler S. 2008. National health spending in 2006: A year of change for prescription drugs. *Health Aff* **27:** 14–29.

Clayton PD, Boebert WE, Defriese GH, Dowell SP, Fennell ML, Frawley KA, Glaser J, Kemmerer RA, Landwehr CE, Rindfleisch TC, et al. 1997. *For the Record: Protecting Electronic Health Information*. National Academy Press, Washington, DC.

Committee on Quality of Health Care in America, Institute of Medicine. 2001. *Crossing the Quality Chasm: A New Health System for the 21st Century*. National Academy Press, Washington, DC.

Cunningham H, Maynard D, Bontcheva K, Tablan V. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*. Association for Computational Linguistics, Philadelphia, PA.

Flynn EA, Barker KN, Pepper GA, Bates DW, Mikeal RL. 2002. Comparison of methods for detecting medication errors in 36 hospitals and skilled-nursing facilities. *Am J Health Syst Pharm* **59:** 436–446.
Girosi F, Meili R, Scoville RP. 2005. *Extrapolating evidence of health information technology savings and costs*. RAND Health, Santa Monica, CA.
Gismondi PM, Hamer DH, Leka LS, Dallal G, Fiatarone Singh MA, Meydani SN. 2005. Strategies, time, and costs associated with the recruitment and enrollment of nursing home residents for a micronutrient supplementation clinical trial. *J Gerontol A Biol Sci Med Sci* **60:** 1469–1474.
Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson JV, Stephan DA, Nelson SF, Craig DW. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet* **4:** e1000167. doi: 10.1371/journal.pgen.1000167.
Ioannidis JP. 2007. Non-replication and inconsistency in the genome-wide association setting. *Hum Hered* **64:** 203–213.
Jollis JG, Ancukiewicz M, DeLong ER, Pryor DB, Muhlbaier LH, Mark DB. 1993. Discordance of databases designed for claims payment versus clinical information systems. Implications for outcomes research. *Ann Intern Med* **119:** 844–850.
Karlawish J, Cary MS, Rubright J, Tenhave T. 2008. How redesigning AD clinical trials might increase study partners' willingness to participate. *Neurology* **71:** 1883–1888.
Kimball R, Ross M. 2002. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*, 2nd ed. John Wiley, New York.
Kohane IS, Mandl KD, Taylor PL, Holm IA, Nigrin DJ, Kunkel LM. 2007. Medicine. Reestablishing the researcher–patient compact. *Science* **316:** 836–837.
Lindberg D, Humphreys B. 1992. The Unified Medical Language System (UMLS) and computer-based patient records. In *Aspects of the computer-based patient record* (eds. M Ball and M Collen), pp. 165–175. Springer-Verlag, New York.
Manolio TA, Bailey-Wilson JE, Collins FS. 2006. Genes, environment and the value of prospective cohort studies. *Nat Rev Genet* **7:** 812–820.
O'Herrin JK, Fost N, Kudsk KA. 2004. Health Insurance Portability Accountability Act (HIPAA) regulations: Effect on medical record research. *Ann Surg* **239:** 772–778.
Ota K, Friedman L, Ashford J, Hernandez B, Penner A, Stepp A, Raam R, Yesavage J. 2006. The Cost–Time Index: A new method for measuring the efficiencies of recruitment-resources in clinical trials. *Contemp Clin Trials* **27:** 494–497.
Poon EG, Jha AK, Christino M, Honour MM, Fernandopulle R, Middleton B, Newhouse J, Leape L, Bates DW, Blumenthal D, et al. 2006. Assessing the level of healthcare information technology adoption in the United States: A snapshot. *BMC Med Inform Decis Mak* **6:** 1. doi: 10.1186/1472-6947-6-1.
Porter SC, Silvia MT, Fleisher GR, Kohane IS, Homer CJ, Mandl KD. 2000. Parents as direct contributors to the medical record: Validation of their electronic input. *Ann Emerg Med* **35:** 346–352.
Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, Masys DR. 2008. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin Pharmacol Ther* **84:** 362–369.
Russell JH, Theodore ES. 2005. Drug records, confidential data vulnerable: Harvard ID numbers, PharmaCare loophole provide wide-ranging access to private data. *The Harvard Crimson*, January 21. http://www.thecrimson.com/article.aspx?ref=505402.
Safran C. 1991. Using routinely collected data for clinical research. *Stat Med* **10:** 559–564.
Safran C, Porter D, Lightfoot J, Rury CD, Underhill LH, Bleich HL, Slack WV. 1989. ClinQuery: A system for online searching of data in a teaching hospital. *Ann Intern Med* **111:** 751–756.
Senate and House of Representatives of the United States of America in Congress. 2009. Title XIII—Health Information Technology. In *Health Information Technology for Economic and Clinical Health Act*, pp. 241–277. http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=111_cong_public_laws&docid=f:publ005.111.pdf. Congress of the USA, Washington, DC.
Spivey A. 2006. Gene-environment studies: Who, how, when, and where? *Environ Health Perspect* **114:** A466–A467.
Sweeney L. 1998. Privacy and medical-records research. *N Engl J Med* **338:** 1078.
Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *J Roy Statist Soc B* **58:** 267–288.
United States Congress Senate Committee on Veterans' Affairs and United States Congress Senate Committee on Homeland Security and Governmental Affairs. 2007. *Veterans Affairs data privacy breach: Twenty-six million people deserve answers: Joint hearing before the Committee on Veterans' Affairs and the Committee on Homeland Security and Governmental Affairs, United States Senate, One Hundred Ninth Congress, second session, May 25, 2006*. U.S. Government Printing Office, Washington, DC.
Warner HR. 1966. The role of computers in medical research. *JAMA* **196:** 944–949.
Wojczynski MK, Tiwari HK. 2008. Definition of phenotype. *Adv Genet* **60:** 75–105.
Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. 2006. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: Evaluation of a natural language processing system. *BMC Med Inform Decis Mak* **6:** 30. doi: 10.1186/1472-6947-6-30.

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