Published in final edited form as:

Sci Transl Med. 2014 April 30; 6(234): 234cm3. doi:10.1126/scitranslmed.3008604.

Biobanks and Electronic Medical Records: Enabling Cost-Effective Research

Erica Bowton^{1,*}, Julie R. Field¹, Sunny Wang¹, Jonathan S. Schildcrout², Sara L. Van Driest³, Jessica T. Delaney⁴, James Cowan¹, Peter Weeke⁴, Jonathan D. Mosley⁴, Quinn S. Wells⁴, Jason H. Karnes⁴, Christian Shaffer⁴, Josh F. Peterson^{4,5}, Joshua C. Denny^{4,5}, Dan M. Roden^{4,6}, and Jill M. Pulley⁷

¹Institute for Clinical and Translational Research, School of Medicine, Vanderbilt University, Nashville, TN 37232, USA

²Department of Biostatistics, School of Medicine, Vanderbilt University, Nashville, TN 37232, USA

³Department of Pediatrics, School of Medicine, Vanderbilt University, Nashville, TN 37232, USA

⁴Department of Medicine, School of Medicine, Vanderbilt University, Nashville, TN 37232, USA

⁵Department of Biomedical Informatics, School of Medicine, Vanderbilt University, Nashville, TN 37232, USA

⁶Department of Clinical Pharmacology, School of Medicine, Vanderbilt University, Nashville, TN 37232, USA

⁷Department of Medical Administration, School of Medicine, Vanderbilt University, Nashville, TN 37232, USA

Abstract

The use of electronic medical record data linked to biological specimens in health care settings is expected to enable cost-effective and rapid genomic analyses. Here, we present a model that highlights potential advantages for genomic discovery and describe the operational infrastructure that facilitated multiple simultaneous discovery efforts.

Traditional studies of drug efficacy and safety address the utility of a specific therapeutic intervention in a defined population. Such study designs present important challenges. Patient accrual can take months to years, and the potential exists for systematic exclusion of clinically complicated but relevant patient groups, such as the elderly, those with comorbid conditions, and those who routinely take multiple drugs. Patient cohorts can be inadequate in size for subgroup analysis, long-term follow-up is often not feasible, and results are limited

*Corresponding author. erica.bowton@vanderbilt.edu.

Supplementary Materials: www.sciencetranslationalmedicine.org/cgi/content/full/6/234/234cm3/DC1

Acknowledgments

Funding

Author contributions

References (28–40)

Competing interests: The authors declare that they have no competing interests.

to diseases for which the participants were originally assessed. Hypothesis-neutral cohorts such as the Framingham Heart Study and Multicenter AIDS Cohort Study (MACS) have overcome these challenges and provided the foundation for critical discoveries that continue to shape health care practice. However, large monetary, time, and infrastructure investments are required to establish and maintain these highly curated, large cohorts in which data collection is focused on hypotheses formulated at the outset.

An alternative to clinical studies with traditional patient cohorts has emerged in the last decade—the pairing of disease-agnostic biobank specimens with electronic medical records (EMRs). Here, we describe the Vanderbilt Electronic Systems for Pharmacogenomic Assessment (VESPA) Project—a large EMR- and biobank-based initiative for translational pharmacogenomic discoveries. We used data from BioVU, Vanderbilt University's EMR-linked biorepository (which as of April 2014 contains more than 179,000 DNA samples) to perform a preliminary cost and time analysis for this approach and compared these costs and time investments with those of traditional cohort studies.

Fashioning an Efficient Pipeline

A key element to establishing an efficient and effective pipeline was the creation of an organizational structure to facilitate communication and management among research teams. Through VESPA, we developed strategies and methods for initiating, executing, and monitoring studies. Essential to this pipeline was the formation of teams for phenotyping and genetic data analysis. Phenotype teams were physician-led and composed of individuals with clinical and informatics expertise, including specific clinical domain content experts. These experts were responsible for cohort selection, algorithm development and refinement, and manual review when necessary. The genetic data-analysis team, which had expertise in laboratory techniques and genomics technologies, directed genotyping assays and interacted with each of the various phenotype teams. Project managers participated in study design, managed both phenotype development and genotyping throughput, and tracked timelines and milestones; this management tier was crucial for promoting multiple, simultaneous studies at different stages of development or execution.

The phenotype pipeline consisted of five key components: selection of a study phenotype, study design, phenotype-specific algorithm development, review, and implementation. Study hypotheses were divided into two categories: (i) validation studies—those that replicated the association of clinical outcomes (for example, drug-response phenotypes) with previously identified genomic variants—and (ii) discovery studies—genome-wide investigations that sought to identify new gene-phenotype associations. A total of 28 phenotypes were selected for study (table S1).

Development of phenotype algorithm

Recent efforts have examined the utility of algorithms for determining phenotypes from EMRs (1–3). We used two approaches to construct phenotype algorithms: (i) fully automated, through the use of phenotype-selection algorithms that achieved high precision, and (ii) semi-automated, using algorithms to select a set of cases for manual review (usually rarer phenotypes). Data sets required to identify cases and controls accurately for each

phenotype varied, but most included three data types: ICD-9 codes, medication regimens, and medical test results. Ten of the phenotypes also required the use of advanced informatics methods, such as natural language processing, to extract information stored in unstructured clinical text.

Pharmacogenomic phenotypes, in particular, rely heavily on temporal relationships (for example, administration of simvastatin before or concurrent with the onset of muscle pain). For our phenotype algorithms, we used event-sequence analyses to establish temporal relationships between drugs and phenotypes, which is a substantial challenge in bioinformatics (4). Both our case and control algorithms excluded records that contained specific clinical comorbidities. Algorithms were quality checked for precision by team members and iteratively refined to achieve positive predictive values (PPV) > 90%. For automated algorithms failing to meet this threshold, manual review was coupled with algorithms to validate that the included cases were true positives (5). Although manual review can be time-consuming and impractical for large cohorts, it is warranted when phenotypes are rare, complex, or involve temporal components too difficult to define electronically.

Enabling overlap

A total of 11,639 subjects (Table 1) met phenotyping criteria for at least one of the 28 phenotypes investigated by the VESPA team. Cohorts included subjects with primarily drug-response phenotypes. Seven phenotypes were not explicitly designed as such but were intended to enable future investigation into potential drug-response phenotypes; for example, subjects exposed to immunosuppressant therapy after organ transplantation offer potential examination of a range of outcomes (drug levels, transplant rejection, lipid abnormalities, cancer, or infections). Across all phenotype cases and controls, 90% were reused as either a case or control for at least one other phenotype. This demonstrates the capability offered by EMR-based studies to reuse cases and controls across both rare and common phenotypes, each with different phenotyping processes. Two VESPA replication studies have established the validity of an EMR-based method for identifying pharmacogenomic associations, clopidogrel major adverse cardiac events, and warfarin stable-dose (5, 6).

Cost Calculations

We compared the estimated monetary cost and resources required to generate VESPA cohorts (excluding analysis) to cost estimates drawn from the analysis of data derived from the NIH RePORTER (7) for M-, R-, U-, P- and Z-type grants that directly supported discrete pharmacogenomics studies in humans. Our analysis (Table 2, legend) revealed striking savings with the multiplexed VESPA approach (Table 2 and Fig. 1). The VESPA experience resulted in 28 case-control sets with a median cost per study of \$76,674 [interquartile range (IQR), \$43,173 to \$207,769] and a median cost per genotyped subject of \$393 (IQR, \$382–\$465). This includes the cost to phenotype cases and controls (personnel resources required to develop algorithms, implement algorithms, extract data, review records, and manage the pipeline) as well as the cost to genotype the cohort (consumables, processing, and quality control).

The median funding amount for pharmacogenomics-related NIH grants with defined cohort sizes (across their lifetimes) is \$1,335,927, with a median cost per genotyped subject of \$1419. Notably, the low median cost per VESPA study (\$76,674) was enabled by the reuse of subjects as cases and controls across multiple studies; had studies been conducted in isolation with no overlap among cases and controls, the estimated median cost per study would have been \$438,473. Further highlighting the efficiency of biobank studies, VESPA studies took a median of 3 months to identify subjects with the target phenotypes, whereas the NIH grants reviewed were awarded for a median period of 3 years. Indeed, traditional consented recruitment models, for example, for common cancers, can take up to 20 years to generate sufficient cohort sizes (8). VESPA studies did not sacrifice cohort size or power as a consequence of reduced cost; in fact, the median cohort size of VESPA phenotypes was 1123, which is almost twice that of NIH-funded pharmacogenomics studies, which had a median cohort size of 623. Compared with a median cost per subject per year of \$478 in a traditional cohort study, the median cost per subject per year in a VESPA study was \$96.

Cost-Saving Infrastructure

There are potential advantages of discovery efforts in an EMR environment, especially when coupled to large genomic resources. First, EMRs contain large patient populations without disease-based exclusions (8). As demonstrated by the EMRs and genomics (eMERGE) network—a U.S. national consortium of existing DNA biorepositories linked to EMRs—these data can be used to rapidly create large, inclusive patient cohorts that foster investigation of variability in physiological traits and disease susceptibility (9–11). Second, the EMR approach offers substantial efficiencies owing to the ability to examine multiple phenotypes by using a single cohort of genotyped samples, an idea first championed on a large scale by the Wellcome Trust Case Control Consortium (12). Third, biobanks enable access, not only to cases but also to large numbers of controls, potentially providing additional power when using a design based on multiple controls per case. Fourth, because EMR-based biobank research is coupled to data routinely obtained in clinical care, the efficiencies of reuse suggest that the approach will prove to be cost-effective. In addition, the increasing use of EMRs [incentivized by the U.S. Health Information Technology for Economic and Clinical Health (HITECH) Act] and the increasing number of EMR-linked biobanks worldwide offer cost-effective resources, not only for discovery but also for the replication of genomic associations across nations and ancestries.

BioVU, the Vanderbilt DNA databank, is an example of an EMR-linked biorepository and a component of eMERGE (13, 14). It is important to note that the total costs described here for the VESPA study are marginal costs—they do not include costs associated with the design, set-up, and building of BioVU or establishing and maintaining the clinical electronic medical record. Thus, the substantial cost savings we observed was facilitated by resources already in place. Development of BioVU, an evolving resource with longitudinal health information, was and is institutionally supported, including investment in EMRs and creation of de-identified images of the EMRs. We highlight the cost savings enabled by BioVU to demonstrate the considerable return on investment afforded by the development of an EMR-based biobank.

As we have demonstrated, EMR-based biobanks can be cost-effective tools for establishing disease or drug associations in a real-world community health care setting. We provide data here that an EMR-linked biobank model such as BioVU enables cost and time efficiency in multiple ways: (i) the use of biological samples that have already been collected and would otherwise be discarded; (ii) an economy of scale obtained by central processing of these samples; (iii) reuse of the same sample for multiple studies without incremental collection, extraction, or processing costs; (iv) centralized de-identification and phenotype annotation of the EMR; and (v) reuse of data, based on program requirements for redeposit of genetic data for all studies. This efficiency is reflected in the substantial cost savings over traditional methods and is further amplified by the ability to examine multiple phenotypes by using a single cohort of genotyped samples (12).

Growth in EMR adoption fostered by the HITECH Act provides the foundation to efficiently expand EMR-based research and is not limited to studies within a single medical center. As evidenced by the robust analyses enabled by the eMERGE network (15–17), the utility of EMR-derived data linked to biological specimens is amplified by pooling analyses across networks, leading to an increase in sample sizes and minimization of biases (18). The eMERGE network has demonstrated successful sharing of more than 18 phenotype algorithms across sites, with a median of three external validations per algorithm. Performance on case and control algorithms for development-site evaluations were similar to external-site evaluations: Median case PPV was 97% for host evaluations, and median PPV for external site evaluations was nearly identical at 95.5%, establishing portability of electronic definitions regardless of the EMR system and interoperability (<http://phekb.org>).

Challenges and Limitations

Data reuse

When combining data from multiple studies in a redeposit design such as that of BioVU, a major challenge is the combining of genotyping data ascertained from different genotyping platforms. This presents challenges for genetic analyses, including the selection of variants for analysis and controlling for batch and platform effects. However, these challenges are not unlike those associated with large genome-wide association study (GWAS) meta-analyses (18–20). Indeed, a key analytical approach for VESPA studies has been to use GWASs, similar to the approach of many traditional pharmacogenomic studies that rely on observational cohorts, subject enrollment, or randomized controlled trials.

Although the GWAS method has been highly successful in identifying new loci associated with disease susceptibility, it has also been criticized because the effect sizes of the identified loci are often small, and thus, very large cohorts are needed to identify and validate genomic variations. On the other hand, although GWAS for drug response traits is less well-explored, multiple studies support the hypothesis that genetic associations can be identified even with small cohort sizes (21–23). Unlike most disease-susceptibility studies, the effect sizes in pharmacogenomics can be large enough to consider for implementation in clinical care. As such, biobanks may become a crucial tool for facilitating pharmacogenomics research. Although we primarily focus on drug-response phenotypes, the

methods described here can be used for a wide range of EMR-derived phenotypes or even to inform phenome-wide analyses (24).

EMR biases

Despite their numerous benefits related to time and efficiency, EMR-linked biobank approaches have limitations (table S2). One fundamental limitation is the potential loss to follow-up or the absence of clinical information pertaining to a patient after a given point in time. In the specific case of BioVU, de-identification of all subjects formally eliminates the ability to recontact patients. Moreover, the data are collected as a result of a provider's determination of need based on clinical relevance at the time and may include only those medical encounters within one given medical center. Thus, studies are limited to, and potentially biased by, data that are available in the EMRs. In addition, it can be challenging to accurately identify cases and controls, particularly for complex phenotypes, and exposure misclassification or selection effect can lead to bias in the estimation of an interaction effect (20, 25).

In our studies, cohorts were defined by an exposure to a medication, a procedure, or patient characteristics at an index point in time; determining cases and controls by temporally constrained definitions can limit cohort populations because of the inherent difficulties in establishing temporality and event sequence in EMR records (26). Moreover, EMR-based data do not inherently capture the cost of a procedure or clinical event. However, an EMR system could be expanded and linked to external data sources, including cost and systems-delivery data, enabling such studies and affording additional opportunities for linking to research-derived data.

Politics

The trend of reduced U.S. federal support for research (27) jeopardizes higher-priced scientific explorations, even those that have proven fruitful for science and health. The current funding climate, rising costs of health care R&D, and stricter payer requirements should make resource reuse increasingly important for advancing clinical and translational research as well as for reducing related health care costs.

The financial efficiencies we observed for the EMR approach make it a compelling complement to traditional cohort designs.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References and Notes

1. Carroll RJ, Eyler AE, Denny JC. Naïve electronic health record phenotype identification for rheumatoid arthritis. *AMIA Annu Symp Proc.* 2011; 2011:189–196. [PubMed: 22195070]
2. Denny JC, Peterson JF, Choma NN, Xu H, Miller RA, Bastarache L, Peterson NB. Extracting timing and status descriptors for colonoscopy testing from electronic medical records. *J Am Med Inform Assoc.* 2010; 17:383–388. [PubMed: 20595304]

3. Carroll RJ, Thompson WK, Eyler AE, Mandelin AM, Cai T, Zink RM, Pacheco JA, Boomershine CS, Lasko TA, Xu H, Karlson EW, Perez RG, Gainer VS, Murphy SN, Ruderman EM, Pope RM, Plenge RM, Kho AN, Liao KP, Denny JC. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *J Am Med Inform Assoc.* 2012; 19(e1):e162–e169. [PubMed: 22374935]
4. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *J Am Med Inform Assoc.* 2013; 20:806–813. [PubMed: 23564629]
5. Delaney JT, Ramirez AH, Bowton E, Pulley JM, Basford MA, Schildcrout JS, Shi Y, Zink R, Oetjens M, Xu H, Cleator JH, Jahangir E, Ritchie MD, Masys DR, Roden DM, Crawford DC, Denny JC. Predicting clopidogrel response using DNA samples linked to an electronic health record. *Clin Pharmacol Ther.* 2012; 91:257–263. [PubMed: 22190063]
6. Ramirez AH, Shi Y, Schildcrout JS, Delaney JT, Xu H, Oetjens MT, Zuvich RL, Basford MA, Bowton E, Jiang M, Speltz P, Zink R, Cowan J, Pulley JM, Ritchie MD, Masys DR, Roden DM, Crawford DC, Denny JC. Predicting warfarin dosage in European-Americans and African-Americans using DNA samples linked to an electronic health record. *Pharmacogenomics.* 2012; 13:407–418. [PubMed: 22329724]
7. RePORT query form. <http://projectreporter.nih.gov/reporter.cfm>.
8. Burton PR, Hansell AL, Fortier I, Manolio TA, Khoury MJ, Little J, Elliott P. Size matters: Just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology. *Int J Epidemiol.* 2009; 38:263–273. [PubMed: 18676414]
9. Kho AN, Pacheco JA, Peissig PL, Rasmussen L, Newton KM, Weston N, Crane PK, Pathak J, Chute CG, Bielinski SJ, Kullo IJ, Li R, Manolio TA, Chisholm RL, Denny JC. Electronic medical records for genetic research: Results of the eMERGE consortium. *Sci Transl Med.* 2011; 3(79re1)
10. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, Li R, Masys DR, Ritchie MD, Roden DM, Struwing JP, Wolf WA. M. E. R. G. E. Team eMERGE Team. The eMERGE network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics.* 2011; 4:13. [PubMed: 21269473]
11. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, Sanderson SC, Kannry J, Zinberg R, Basford MA, Brilliant M, Carey DJ, Chisholm RL, Chute CG, Connolly JJ, Crosslin D, Denny JC, Gallego CJ, Haines JL, Hakonarson H, Harley J, Jarvik GP, Kohane I, Kullo IJ, Larson EB, McCarty C, Ritchie MD, Roden DM, Smith ME, Böttiger EP, Williams MS. eMERGE Network. The electronic medical records and genomics (eMERGE) network: past, present, and future The electronic medical records and genomics (eMERGE) network: past, present, and future. *Genet Med.* 2013; 15:761–771. [PubMed: 23743551]
12. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls. *Nature.* 2007; 447:661–678. [PubMed: 17554300]
13. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, Masys DR. Development of a largescale de-identified DNA biobank to enable personalized medicine. *Clin Pharmacol Ther.* 2008; 84:362–369. [PubMed: 18500243]
14. McGregor TL, Van Driest SL, Brothers KB, Bowton EA, Muglia LJ, Roden DM. Inclusion of pediatric samples in an opt-out biorepository linking DNA to de-identified medical records: Pediatric BioVU. *Clin Pharmacol Ther.* 2013; 93:204–211. [PubMed: 23281421]
15. Ritchie MD, Denny JC, Zuvich RL, Crawford DC, Schildcrout JS, Bastarache L, Ramirez AH, Mosley JD, Pulley JM, Basford MA, Bradford Y, Rasmussen LV, Pathak J, Chute CG, Kullo IJ, McCarty CA, Chisholm RL, Kho AN, Carlson CS, Larson EB, Jarvik GP, Sotoodehnia N, Manolio TA, Li R, Masys DR, Haines JL, Roden DM. Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) QRS Group. Genome-and phenome-wide analyses of cardiac conduction identifies markers of arrhythmia risk. *Circulation.* 2013; 127:1377–1385. [PubMed: 23463857]
16. Kullo IJ, Ding K, Shameer K, McCarty CA, Jarvik GP, Denny JC, Ritchie MD, Ye Z, Crosslin DR, Chisholm RL, Manolio TA, Chute CG. Complement receptor 1 gene variants are associated with erythrocyte sedimentation rate. *Am J Hum Genet.* 2011; 89:131–138. [PubMed: 21700265]
17. Denny JC, Ritchie MD, Crawford DC, Schildcrout JS, Ramirez AH, Pulley JM, Basford MA, Masys DR, Haines JL, Roden DM. Identification of genomic predictors of atrioventricular

- conduction: Using electronic medical records as a tool for genome science. *Circulation*. 2010; 122:2016–2021. [PubMed: 21041692]
18. Ioannidis JPA, Trikalinos TA, Khoury MJ. Implications of small effect sizes of individual genetic variants on the design and interpretation of genetic association studies of complex diseases. *Am J Epidemiol*. 2006; 164:609–614. [PubMed: 16893921]
 19. Evangelou E, Ioannidis JPA. Meta-analysis methods for genome-wide association studies and beyond. *Nat Rev Genet*. 2013; 14:379–389. [PubMed: 23657481]
 20. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, Hirschhorn JN. Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nat Rev Genet*. 2008; 9:356–369. [PubMed: 18398418]
 21. Cooper GM, Johnson JA, Langae TY, Feng H, Stanaway IB, Schwarz UI, Ritchie MD, Stein CM, Roden DM, Smith JD, Veenstra DL, Rettie AE, Rieder MJ. A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. *Blood*. 2008; 112:1022–1027. [PubMed: 18535201]
 22. Link E, Parish S, Armitage J, Bowman L, Heath S, Matsuda F, Gut I, Lathrop M, Collins R. SEARCH Collaborative Group. SLCO1B1 variants and statin-induced myopathy—A genome-wide study. *N Engl J Med*. 2008; 359:789–799. [PubMed: 18650507]
 23. Mallal S, Phillips E, Carosi G, Molina JM, Workman C, Tomazic J, Jägel-Guedes E, Rugina S, Kozyrev O, Cid JF, Hay P, Nolan D, Hughes S, Hughes A, Ryan S, Fitch N, Thorborn D, Benbow A. PREDICT-1 Study Team. HLA-B*5701 screening for hypersensitivity to abacavir. *N Engl J Med*. 2008; 358:568–579. [PubMed: 18256392]
 24. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, Field JR, Pulley JM, Ramirez AH, Bowton E, Basford MA, Carrell DS, Peissig PL, Kho AN, Pacheco JA, Rasmussen LV, Crosslin DR, Crane PK, Pathak J, Bielinski SJ, Pendergrass SA, Xu H, Hindorff LA, Li R, Manolio TA, Chute CG, Chisholm RL, Larson EB, Jarvik GP, Brilliant MH, McCarty CA, Kullo IJ, Haines JL, Crawford DC, Masys DR, Roden DM. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol*. 2013; 31:1102–1110. [PubMed: 24270849]
 25. Garcia-Closas M, Rothman N, Lubin J. Misclassification in case-control studies of gene-environment interactions: Assessment of bias and sample size. *Cancer Epidemiol Biomarkers Prev*. 1999; 8:1043–1050. [PubMed: 10613335]
 26. Jiang M, Denny JC, Tang B, Cao H, Xu H. Extracting semantic lexicons from discharge summaries using machine learning and the C-Value method. *AMIA Annu Symp Proc*. 2012; 2012:409–416. [PubMed: 23304311]
 27. The impact of sequestration on NIH. 2012. www.aamc.org/research/adhocgp/aamcimpactofsequestrationonnih.pdf
 28. Collins FS. The case for a US prospective cohort study of genes and environment. *Nature*. 2004; 429:475–477. [PubMed: 15164074]
 29. Kohane IS. Using electronic health records to drive discovery in disease genomics. *Nat Rev Genet*. 2011; 12:417–428. [PubMed: 21587298]
 30. Henderson GE, Cadigan RJ, Edwards TP, Conlon I, Nelson AG, Evans JP, Davis AM, Zimmer C, Weiner BJ. Characterizing biobank organizations in the U.S.: Results from a national survey. *Genome Med*. 2013; 5:3. [PubMed: 23351549]
 31. Ollier W, Sprosen T, Peakman T. UK Biobank: From concept to reality. *Pharmacogenomics*. 2005; 6:639–646. [PubMed: 16143003]
 32. Palmer LJ. UK Biobank: bank on it. *Lancet*. 2007; 369:1980–a1982. [PubMed: 17574079]
 33. Chen Z, Chen J, Collins R, Guo Y, Peto R, Wu F, Li L. China Kadoorie Biobank (CKB) collaborative group. China Kadoorie Biobank of 0.5 million people: Survey methods, baseline characteristics and long-term follow-up. *Int J Epidemiol*. 2011; 40:1652–1666. [PubMed: 22158673]
 34. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: A medication information extraction system for clinical narratives. *J Am Med Inform Assoc*. 2010; 17:19–24. [PubMed: 20064797]

35. Trinidad SB, Fullerton SM, Bares JM, Jarvik GP, Larson EB, Burke W. Genomic research and wide data sharing: Views of prospective participants. *Genet Med*. 2010; 12:486–495. [PubMed: 20535021]
36. Brothers KB, Clayton EW. Parental perspectives on a pediatric human non-subjects biobank. *AJOB Prim Res*. 2012; 3:21–29. [PubMed: 23181193]
37. Pulley JM, Brace MM, Bernard GR, Masys DR. Attitudes and perceptions of patients towards methods of establishing a DNA biobank. *Cell Tissue Bank*. 2008; 9:55–65. [PubMed: 17960495]
38. Simon CM, Newbury E, L'heureux J. Protecting participants, promoting progress: Public perspectives on community advisory boards (CABs) in biobanking. *J Empir Res Hum Res Ethics*. 2011; 6:19–30. [PubMed: 21931234]
39. Murphy J, Scott J, Kaufman D, Geller G, LeRoy L, Hudson K. Public perspectives on informed consent for biobanking. *Am J Public Health*. 2009; 99:2128–2134. [PubMed: 19833988]
40. Scott CT, Caulfield T, Borgelt E, Illes J. Personal medicine—The new banking crisis. *Nat Biotechnol*. 2012; 30:141–147. [PubMed: 22318029]

Fig. 1. Time is money

Comparison of traditional NIH-funded pharmacogenomic studies versus EMR/biobank studies (BioVU). **(Left)** Median cost of study per subject. **(Right)** Median length of study in years.

Table 1
VESPA cohorts and phenotypes

Total number of genotyped subjects: 11,639
Total number of phenotypes analyzed: 28*
Median age: 61.6 years (range, newborn to 100+)
Observer-reported race: ~84% Caucasian, 12% African American
Subject phenotypic data: Majority had medical records with rich phenotypic data (median of 80 diagnosis codes and a median of 7.7 years of follow-up, from the first to last electronic clinical note)
Median cohort size [†] : 1123 (IQR, 492 to 4158)
Median case cohort size [‡] : 133 (IQR, 84 to 569)
Total case counts [‡] : Ranged from 6 total cases (cerebrovascular event following clopidogrel therapy) to 1174 total cases (cough attributed to ACE inhibitor exposure)
Genomic data available: <ul style="list-style-type: none"> • Genome-wide genotyping data were already available in 2500 subjects • 9139 subjects were newly genotyped in both GWAS and drug-metabolism platforms • An additional 693 subjects and 1167 subjects previously underwent candidate SNP genotyping for clopidogrel adverse events or warfarin stable dose, respectively (5, 6)

* Clopidogrel in cardiovascular disease, warfarin stable dose, early repolarization, vancomycin, C. difficile colitis, anthracycline cardiomyopathy, Guillain-Barre Syndrome, heart transplant, kidney transplant, clopidogrel in cerebrovascular disease, statin-related myopathy, heparin-induced thrombocytopenia, cardiovascular events during COX2 inhibition therapy, serious bleeding during warfarin therapy, amiodarone toxicity (lung, thyroid), chronic inflammatory polyneuropathy, rheumatic heart disease, cough during ACE inhibitor therapy, fluoroquinolones and tendonitis/tendon rupture, warfarin stable dose in children, metformin efficacy, metformin and cancer survival, bisphosphonates and atypical fracture/jaw osteonecrosis, Wolff-Parkinson-White, steroid-induced osteonecrosis, shellfish anaphylaxis, aspirin anaphylaxis, and Bell's Palsy.

[†] Cases and controls.

[‡] Additional phenotype counts are shown in table S1.

Table 2
NIH-funded pharmacogenomic versus EMR-biobank studies*

	Traditional study	BioVU study
Median cohort size (IQR)	623 (273 to 2095)	1123 (492 to 4158)
Median reuse of cohort (IQR)	N/A	55% (34 to 98%)
Median cost (in U.S. dollars) (IQR)	\$1,335,927 (\$416,895 to \$2,715,895)	\$76,674 (\$43,173 to \$207,769)
Median cost/subject (IQR)	\$1419 (\$456 to \$4672)	\$393 (\$382-\$465)
Median years of study (IQR)	3 (2 to 5)	0.25 (0.17 to 0.56)
Median cost/yr/subject (IQR)	\$478 (\$134 to \$1216)	\$96 (\$55 to \$194)

* Funding data for traditional human pharmacogenomic studies were obtained by querying NIHReporter for all funded M-, R-, U-, P- and Z-type grants that contained the keywords “pharmacogenetic” or “pharmacogenomic” (query performed on 2 November 2012). The resulting grant abstracts were reviewed manually to ensure that they directly supported human pharmacogenomics research and to identify the number of subjects in the proposed study cohort. Excluded were studies with only in vitro or animal-model experiments, those directed solely at technology development, and those for which a defined study-cohort size or clinical trial protocol could not be determined. Dollars awarded and years of the award to date were summed for 115 unique NIH grants. Cohort size (total cases plus controls), cost, and time-investment data for VESPA phenotypes were recorded internally. For each phenotype, time investment was calculated as the amount of time required to develop and implement phenotype algorithms, extract data, review records, and complete phenotype curation. Total cost of the VESPA study was calculated on the basis of the number of hours invested and the hourly rate of personnel required to complete the phenotyping plus the cost of genotyping the cohort.