Surveying Recent Themes in Translational Bioinformatics: Big Data in EHRs, Omics for Drugs, and Personal Genomics

J. C. Denny
Vanderbilt University Medical Center, Departments of Biomedical Informatics and Medicine, Nashville, TN, USA

Summary

Objective: To provide a survey of recent progress in the use of large-scale biologic data to impact clinical care, and the impact the reuse of electronic health record data has made in genomic discovery.

Method: Survey of key themes in translational bioinformatics, primarily from 2012 and 2013.

Result: This survey focuses on four major themes: the growing use of Electronic Health Records (EHRs) as a source for genomic discovery, adoption of genomics and pharmacogenomics in clinical practice, the possible use of genomic technologies for drug repurposing, and the use of personal genomics to guide care.

Conclusion: Reuse of abundant clinical data for research is speeding discovery, and implementation of genomic data into clinical medicine is impacting care with new classes of data rarely used previously in medicine.

Keywords

Translational bioinformatics, personalized medicine, electronic health records, genomics, big data

Yearb Med Inform 2014:199-205
http://dx.doi.org/10.15265/IY-2014-0015
Published online August 15, 2014

Introduction

Since the completion of the first draft of the human genome project more than a decade ago, much has been learned about the structure of the human genome and the genomic architecture underlying common and rare human diseases and traits [1, 2]. Propelled by new sequencing technologies and massive populations for large-scale genetic studies, our understanding of the influence of genetics in rare and complex human traits has been increasing, yielding insights into thousands of diseases and trait [3, 4]. Translation of these discoveries in terms of clinical impact are more recent phenomena and the specific focus of the field of translational bioinformatics (TBI). This paper is a survey of several themes in TBI over the last several years.

As defined by Altman and Miller [5], TBI can be defined as the integration of basic molecular, genetic, or cellular data with clinical data for discovery or implementation. Summarizing recent developments in a field as diverse and large as TBI can be challenging. As a result, I first surveyed a number of TBI articles published primarily in 2012 and 2013 to generate a list of recent themes. Impactful papers relevant to TBI were "nominated" via four methods: papers nominated for or highlighted in Dr. Altman's TBI "Year in Review" keynote presentations presented at the 2012 and 2013 AMIA Summit on TBI (slides available at http://rbaltman.wordpress.com), nomination by one of ~20 noted investigators informally surveyed by email (they were asked for papers published in 2012 and 2013), and by my personal review of the field, including review of the submissions to the 2014 AMIA Summit on TBI. Thus, this survey doesn't include all themes in TBI nor is systematic, and certainly omits a number of worthy papers and resources that could fall within the domain of TBI.

After reviewing these papers, I selected four major themes. In two of these themes, Electronic Health Records (EHRs) play a key role. The first was the use of EHRs as a source for genomic discovery through the role of EHR-linked biobanks. The same techniques applied to EHRs are enabling other uses of EHRs for discovery as a source of clinical "Big Data." The second theme is the adoption of genomics and pharmacogenomics as part of "routine" clinical care, in which EHRs also play a critical role. This theme was found in the 2013 AMIA Fall Symposium and 2014 AMIA Summit on TBI among presentations, panels, and a "Birds of a Feather" session. The third theme is the applications of omic technologies to feed drug discovery and repurposing. This aim includes seminal works that highlight the potential for these basic discoveries to impact clinical care. Finally, the fourth theme follows recent trends in personal genomic testing to guide current and future clinical care, including the changing landscape in Direct-To-Consumer (DTC) genetic testing and recent legal challenges regarding *BRCA1* and *BRCA2* testing. This paper focuses primarily on papers published in 2012 and 2013, but selected papers before 2012 have also been included based on their particular relevance to these themes, to provide background, or to credit an idea or method.

Electronic Health Records as Big Data for Genomic Discovery

Use of electronic health record (EHR) data as a source for genomic discovery is a recent development but proving to be a powerful tool to investigate the genetic basis of disease and drug response. Following completion of the first human genome sequence in 2003, the first genome-wide association study (GWAS) was performed in 2005 and, by 2010, the count had reached 500 and covered a wide range of human diseases and traits [1]. Recognizing the potential of EHRs as a tool for genomic medicine, the National Human Genome Research Institute (NHGRI) established the Electronic Medical Records and Genomics (eMERGE) Network in 2007 [6], which was renewed in 2011 and currently includes 10 sites with EHRs linked to DNA biobanks [7]. The first EHR-based genomic studies were published in 2010 [8-11]. At the end of 2013, sites from the eMERGE network alone have published 35 studies using EHR data for genetic analysis. The work of eMERGE has highlighted the importance of an iterative creation-evaluation-refinement process to build "phenotype algorithms", typically composed of elements from billing codes, medication data, laboratory and test results, and/or natural language processing applied to clinical documentation [12]. Many of these algorithms are available on http://phekb.org. Initial studies in the eMERGE network proceeded using data from a single site but the power of cooperative, multi-site studies was quickly realized and accomplished by repurposing existing genetic data with shared phenotype algorithms to identify more cases or to even perform entirely *in silico* genetic studies [13]. Indeed, this model was adopted for eMERGE-II, in which all GWAS performed use extant genotypes.(7) Efforts within eMERGE [14] and SHARPn [15] have highlighted the possibility of creating standardized, computable phenotypes; for now, these phenotypes algorithms are largely shared as descriptive, human-readable documents.

Pharmacogenetic studies have also been published using EHR data [16-18]. EHRs may constitute an important platform for pharmacogenetics research in that they provide a longitudinal collection of medication and disease exposures, and may allow for capture of rare and potentially fatal events when linked to prospective biobanks [19]. They may also be significantly faster, cheaper, and have the ability to accrue more patients than traditional methods of performing pharmacogenomic studies [20]. However, pharmacogenetic studies using EHR data can be complicated by incomplete knowledge of patient adherence, inaccurate medication start and stop dates, and the need to sequence multiple diseases and exposures to accurately assess a pharmacogenetic phenotype.

EHRs have also provided the ability to analyze many diverse phenotypes. Phenome-wide association studies (PheWAS) provide a systematic approach to analyze phenotypes associated with a given genotype, and demonstrate a paradigm shift toward looking at the phenome in a hypothesis-free manner, much as genomic investigation has done [21]. PheWAS often uses aggregations of EHR billing codes to define cases and controls [11]. In the last two years, much larger studies have been performed to follow-up on GWAS discoveries and to analyze specific genetic variants of interest. PheWAS been performed within EHR data sets focused on genetic variants of interest [22] and also on follow up on GWAS results [13, 23, 24]. PheWAS in conjunction with GWAS has noted associations not apparent in the GWAS; e.g., future development of atrial fibrillation in patients with variants associated with slower ventricular conduction [23]. A PheWAS of 3,144 SNPs with prior GWAS associations using 13,835 individuals in the eMERGE network replicated 66% of sufficiently-powered known GWAS associations and found 63 new, potentially pleiotropic, associations, the strongest of which were replicated in another population [25]. The PheWAS approach has also been applied to non-EHR, population-based studies. In a study of 70,061 study participants in the Population Architecture using Genomics and Epidemiology (PAGE) network, Pendergrass et al. studied 83 SNPs across 105 phenotype classes and identified 111 SNP-phenotype associations that passed a nominal significance threshold in at least two sites; among these, 33 represented potentially novel findings [26]. Available tools should enable broader access to PheWAS in EHR and non-EHR datasets [25, 27].

Other large initiatives will dramatically increase the number of patients available in EHR-linked biobanks. The Veteran Affairs (VA) Million Veteran Program (MVP) aims to collect DNA from one million veterans treated at VA hospitals [28]. The primary source of phenotypic information for MVP derives from the VA's extensive national EHR system that links all VA hospitals, clinics, and pharmacies. The Kaiser Permanente Research Program on Genes, Environment, and Health seeks to collect DNA on 500,000 individuals [29], and currently includes more than 100,000 individuals with genome-wide genetic data. The UK Biobank [30] and China Kadoorie biobank [31] have each created population-based longitudinal cohorts with over 500,000 individuals, each with detailed prospective questionnaires, sample collections, focused testings, and the potential for recontact. Importantly, these national biobanks are integrating EHR data from patients into their data repositories.

Use of Electronic Health Record Data for Non-genomic Discovery

Genomic data is clearly not the only application that has benefited from the secondary use of large-scale EHR data. LePendu et al. used natural language processing to mine full-text clinical notes for detecting drug–adverse event associations and for detecting drug–drug interactions [32]. A similar analysis was performed in the Stanford clinical data warehouse to show that cilostazol, the only medication approved for use in peripheral vascular disease but which carries a black box warning for cardiovascular mortality, was not associated with increased adverse cardiac events or overall mortality [33]. Ryan et al. performed perhaps the first "medication-wide association study" to analyze the association between a broad range of medications both individually and by drug class on four clinical outcomes; they replicated a number of known drug-adverse event pairs and suggested new associations, though challenges remain in interpretation of comorbid disease indications and non-random polypharmacy, both of which

can confound results [34]. A study in 2011 used a free-text query-facilitated review of EHR data to determine whether to give or not anticoagulants to a pediatric patient with lupus in the absence of published evidence [35]. Collectively, these studies point to a future use of the EHR as a source of "big data" to guide care, potentially as a real-time consult for patient care.

Similarly, phenome-wide analyses of EHR data are not limited to genetic correlations. Liao et al. used PheWAS to analyze associations between disease phenotypes and autoantibodies [36]. Neuraz et al. extended PheWAS to map ICD-10 codes to correlate thiopurine methyltransferase (TPMT) activity with phenotypes, noting that those with increased TPMT activity were more likely to have outcomes associated with inadequate treatment with TPMT inhibitors [37]. Boland et al. used a PheWAS method to evaluate periodontal disease, noting associations with diabetes, hypertension, and hypercholesterolemia [38]. Doshi-Velez et al. found autism spectrum disorder patients could be grouped by their clusters of comorbid disease [39] Analysis of such clinical data may inform segregation of diseases into subtypes, informative for biologic analysis. In an analysis of more than 110 million patients from the US and Denmark, Blair et al. used billing code data to show that complex diseases often co-occurred with Mendelian diseases, replicating known Mendelian-complex diseases associations (e.g., ataxia telangiectasia and breast cancer) and suggesting many new ones, such as Fragile X with several autoimmune diseases [40]. In support of this finding, they found Mendelian genes were overrepresented among genes associated with common diseases through GWAS.

The growth of secondary use of EHRs for clinical, genomic, and pharmacogenomic research (as well as a future promise of use for other omic technologies) calls into question the nature of the EHR itself. As Hripcsak and Albers argue [41], while the EHR provides an unprecedented resource of longitudinal, detailed clinical data, it may be incomplete [42], fragmented [43], and erroneous [10, 25] at times. Thus, new methods of studying the EHR as an "object of interest in itself" [41] are needed as we consider the EHR as an active participant in the process of phenotyping.

Adoption of Genomics and Pharmacogenomics in Clinical Practice

The promise of omic technologies has always been in their application to clinical care. Francis Collins, the current Director of the National Institutes of Health, recognized that implementation of genomic medicine required preemptive genotyping and the use of EHRs to automate the process, saying, in 2009, that "if everybody's DNA sequence is already in their medical record and it is simply a click of the mouse to found out all the information you need, then there is going to be a much lower barrier to beginning to incorporate that information into drug prescribing." [44] Implementation of this vision requires detailed genomic (or other large-scale biologic data) be available, interpretable, and made clearly actionable to a broad range of clinicians. A number of institutions have since been exploring pragmatic implementations of genomic medicine projects and the necessary role the EHR plays in adoption.

Sarkar identified genomic medicine clinical implementation efforts as major developments in his 2012 IMIA Yearbook Survey [45], and significantly more progress has been made over the last year. Sarkar noted two programs in 2012: the Vanderbilt Pharmacogenomic Resource for Enhanced Decisions in Care & Treatment (PREDICT) [46] and a similar effort targeted for the pediatric cancer population at St. Jude Children's Research Hospital [47]. Both of these employ multiplexed genotyping assays to evaluate common pharmacokinetic and pharmacodynamics variants and provide raw and interpreted genetic results within the EHR. Analysis of the first ~10,000 patients in PREDICT noted that 91% of the European ancestry patients and in 96% of African ancestry patients carried an actionable genetic variant for at least one of the five implement drug-genome interactions [48]. Moreover, preemptive multiplexed genetic testing resulted in 46% fewer tests performed when compared with a reactive strategy that tested when each target medication was prescribed. The 1200 Patients Project at the University of Chicago will recruit 1200 patients from 12 pre-selected physicians for prospective genetic testing [49]. Information on genetic variants is provided through a custom web interface that displays summarized phenotype information. The University of Florida and Shands Hospital's Personalized Medicine Program is testing patients undergoing cardiac catheterization with a custom array of 256 SNPs with the goal of evaluating the effect of clinical alerts on clopidogrel prescribing [50]. Other eMERGE sites have also engaged in clinical implementation projects, including the CLIPMERGE-PGx project at Mount Sinai [51], Geisinger Health System, Mayo clinic, and others. Common features of many of these implementations are the inclusion of select variants in the EHR with clear actionability (as opposed to a broad set of genotype results of unknown significance), phenotype interpretation, and oversight by institutional Pharmacy and Therapeutics committees (see Figure 1). The recent Implementing Genomics in Practice (IGNITE) network funded by NHGRI seeks to integrate genomic information into EHRs and develop genomic clinical decision support at sites beyond large academic hospitals [52].

An important component of implementation of pharmacogenetics in practice is clear guidance on what to do once an actionable variant is noted. An important effort in this field has been the Clinical Pharmacogenetics Implementation Consortium (CPIC), a shared project of PharmGKB (http://pharmgkb.org) and the Pharmacogenetics Research Network (PGRN) [53]. CPIC guidelines have been published for 23 medications to date [54]. Indeed, in the case of clopidogrel [55, 56], the guideline has been revised to incorporate new evidence, reflecting the fast-moving nature of pharmacogenomic knowledge. The FDA lists over 100 medications with germline variants that could affect drug prescribing [57].

Using Omic Technologies for Drug Discovery and Repurposing

The cost of generating new therapeutics has risen dramatically over the past 60 years, with each new drug costing about 80-fold more in 2010 than 1960 in inflation-adjusted terms [58]. As a result, many are investigating

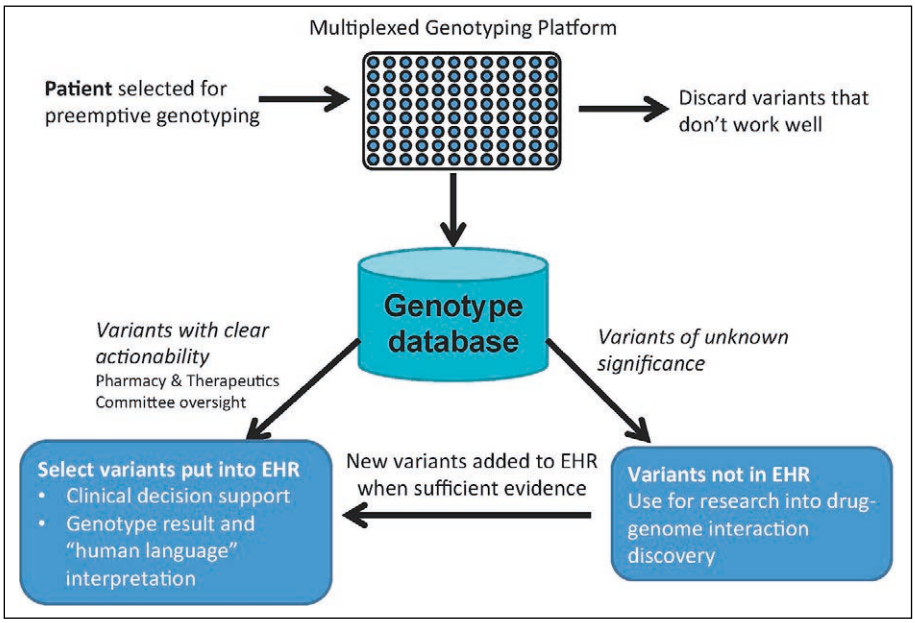

Fig. 1 Model for Clinical Implementation of Pharmacogenomics. This model is built off implementations used by PREDICT(46), CLIPMERGE(51), and University of Florida(50).

high-throughput and computational approaches to drug discovery and repurposing. Recent efforts have focused on the use of the omics data, especially genomics, to discover new drug targets and search for new uses for existing drugs, referred to as drug repositioning. In two linked papers, Dudley et al. [59] and Sirota et al.[60] created disease signatures from microarray data in Gene Expression Omnibus and compared these to gene expression data from Connectivity Map [61] to identify potentially novel therapeutics for lung cancer and inflammatory bowel disease. A similar study using this method, noted that tricyclic antidepressants may have efficacy against small cell lung cancer (but not non-small cell lung cancer) [62].

Disease-gene association data may also predict drug targets. Sanseau et al. evaluated existing GWAS hits and found that genes related to GWAS hits are significantly more likely to be targetable by small molecules or by biologic agents than other genomic regions, and that 15.6% of GWAS genes are existing drug targets (compared to 5.7% of the general genome) [63]. In support of this hypothesis, Okada et al. performed a multi-ethnic GWAS of 103,638 cases and controls for rheumatoid arthritis (RA) and noted 101 total RA risk loci; these loci identified 18 of 27 current RA drug target genes, and identified three approved cancer medications that may be active against RA [64]. Khatri et al. analyzed eight existing organ transplant rejection datasets and found a common module of 11 genes overexpressed in all rejected organs [65]. Using these genes, they identified two existing non-immunosuppressant drugs that could be repurposed to regulate these genes, and demonstrated enhanced effect in a mouse model. Sateriale et al. analyzed protozoan genomes to predict antimicrobial activity, then validated predicted activity against one protozoan with prior results from a cell-based assay [66]. Resources such as the drug-gene interaction database (http://DGIdb.org), which integrates data from 13 databases [67], and PharmGKB (http://pharmgkb.org) may facilitate translation of genomic study results to potential therapeutics. See the Table for a listing of TBI resources discussed in this paper.

Trends in Personal Genomic Testing to Guide Health Care

Direct-To-Consumer (DTC) genetic testing through sites such as 23andMe (Mountain View, CA) has provided an avenue for patients to pursue genetic testing outside of a doctor's order. Individuals received test results and personalized information on their genetic ancestry, disease risk, and drug response for selected medications. Tenenbaum et al. described a model for how DTC genetic testing could be used to guide care with clinical input [68]. They reported the case of a woman with unremarkable personal and family history who learned through DTC testing about the presence of a prothrombin gene mutation, and as a result, underwent anticoagulation during pregnancy. Each DTC website maintained its own predictions of disease risk by curating genetic risk variants and effect sizes from literature evidence. As a result, risk predictions can vary, incorporating different SNPs and resulting in varying classification performance using simulated population data [69]. In addition to personal information, 23andMe has also been an important force in genomic discovery. Using electronic surveys completed by 23andMe enrollees, they replicated over 180 known GWAS associations, including 75% of those for which they adequately powered [70], and have discovered new associations for other diseases and traits, such as environmental allergies [71] and hypothyroidism [72].

Recent events have made clinical guidance from DTC genetic testing harder to obtain for new customers. Navigenics and deCODEme, two other sites that previously provided DTC genetic testing, have withdrawn from the personal genome-testing market after being purchased by Life Technologies Corp. (Grand Island, NY) and Amgen Inc. (Thousand Oaks, CA), respectively. Furthermore, on November 22, 2013, the Food and Drug Administration ordered 23andMe to stop providing clinical guidance for genetic test results, citing "potential health consequences that could result from false positive or false negative assessments." [73] As a result, 23andMe is no longer providing disease risk and drug response information to new enrollees, though such information remains available to prior enrollees at the time of this writing.

Another major development in 2013 regarding genetic testing came in the form of the Supreme Court case *Association for Molecular Pathology v. Myriad Genetics*, which ruled that genes or specific naturally-occurring variants cannot be patented, opening the possibility for other labs to test

for *BRCA1* and *BRCA2* variants, specifically, and more broadly ensuring the ability to test for actionable variants for other genes.

Dense biologic data may have an emerging role in individual healthcare. In a landmark study, Chen et al. performed genomic, transcriptomic, proteomic, metabolomic, and autoantibody profiles from a single individual over a 14 month period [74]. Genomic testing indicating increased risk of type 2 diabetes, and regular metabolic profiling noted transient elevation hemoglobin A1c levels to a diabetes-defining 6.7% following a respiratory syncytial virus infection. Sequencing technologies may see their first large-scale role in defining somatic mutations in cancers. Candidate SNP genotyping programs for melanoma can influence care choices and enrollment into clinical trials [75]. Next generation sequencing has revealed new mutations in known causative genes, potentially expanding both testing and treatment indications [76], and have revealed successful treatments for existing cancers not previously envisioned [77, 78]. Large-scale analysis of multiple omic data platforms from The Cancer Genome Atlas suggests a future classification of cancer, not by tissue, but by mutational analysis [79, 80].

Sequencing data may also have a role in tracking infections. Snitkin et al. used whole genome sequencing to track an outbreak of a resistant Klebsiella infection at the NIH clinical center to a single patient [81]. Sequencing also helped identify that the reemergence of cholera in Haiti in 2010 as from a foreign source [82], and characterize that there were two distinct populations at the outset of the disease [83].

Looking Forward

EHRs and EHR-linked biobanks are rapidly becoming a very valuable source of big data for discovery. To reach their true potential, systems to link them together and rapidly execute common phenotypes across very large populations are needed. Significant clinical informatics challenges remain to execute this vision. The challenge for the next decade of genomics is translation of large-scale biological data-driven discovery into clinical impact. To do so, EHRs will need to easily manipulate big data, and not just be a generator of it. The first large-scale biologic data imported will likely be genomic. The data will likely be segregated into actionable and non-actionable variants, as suggested by ongoing clinical genomic implementation efforts mentioned above. Although "non-actionable" content does not need to be immediately available to clinicians, it needs to be stored, since the data may become relevant with future discoveries [84]. A new breed of clinical decision support systems (CDS) is needed to easily guide providers to clinical interpretations of dense genomic information, which often have non-intuitive nomenclatures. CDS systems must also be able to be changed quickly with evolving evidence [55, 56]. Clinical adoption will be facilitated by consensus national guidelines (such as CPIC) that start from assumption that dense genetic data are already embedded in the EHR. The growth of secondary use of EHR data within sites and aggregated across networks of institutions will also play a key role in discovery, as future incorporation of dense genetic data into EHRs will enable to add new classes of clinical discovery. Given genetic differences in ancestral populations, accrual of diverse populations will be critical to develop evidence to guide and refine care over time, with the ultimate goal of an omic-enabled, learning healthcare system.

Table1 Select public resources available for Translational Bioinformatics.

Name	URL	Comments
Pharmacogenomic Biomarkers in Drug Labels	http://www.fda.gov/drugs/scienceresearch/ researchareas/ pharmacogenetics/ucm083378.htm	Lists FDA-approved drugs with pharmacogenomic information in their drug labels.
PharmGKB	http://www.pharmgkb.org	PharmGKB is a curated resource about the impact of genetic variation on drug response for clinicians and researchers.
Clinical Pharmacogenetics Implementation Consortium (CPIC)	http://www.pharmgkb.org/page/cpic	Provides a list of the published guidelines for drug-gene interactions produced by CPIC.
Phenotype Knowledgebase	http://phekb.org	Online collaborative repository for building, validating, and sharing electronic phenotype algorithms and their performance characteristics.
NHGRI Catalog of GWAS studies	http://www.genome.gov/26525384	Curated list of GWAS studies, their phenotypes, and key results.
Catalog of PheWAS results	http://phewascatalog.org	Searchable, downloadable catalog of EHR PheWAS results.
Drug-Gene Interaction database	http://dgidb.genome.wustl.edu	Provides a search interface into drug-gene interactions from data derived from 13 resources.
My Cancer Genome	http://www.mycancergenome.org	Provides up-to-date data regarding cancer mutations, treatments, and relevant clinical trials.
ClinVar	http://www.ncbi.nlm.nih.gov/clinvar/	It provides up-do-date relationships among human variations and phenotypes along with supporting evidence.
SHARPn	http://phenotypeportal.org	Collection of computable phenotype algorithms generated by SHARPn.

Conflict of Interest Notification

The author declares no conflict of interest.

Acknowledgements

The author wishes to thank Dr. Russ Altman for his previous TBI Year in Review presentations and permission to use his work to inform this one, and the help of the corresponding investigators who provided suggestions. There are certainly other seminal works in TBI omitted from this review that should have been included, and such omissions are solely the author's responsibility. Sources of funding supporting this work include funding from National Human Genome Research Institute (U01 HG006378) and the National Library of Medicine (R01 LM010685). Its contents

are solely the responsibility of the author and do not represent official views of the National Institutes of Health.

References

1. Green ED, Guyer MS. Charting a course for genomic medicine from base pairs to bedside. Nature 2011 Feb 10;470(7333):204–13.
2. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature 2004 Oct 21;431(7011):931–45.
3. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 2009 Jun 9;106(23):9362–7.
4. OMIM - Online Mendelian Inheritance in Man [Internet]. [cited 2014 May 20]. Available from: http://omim.org/
5. Altman RB, Miller KS. 2010 Translational bioinformatics year in review. J Am Med Inform Assoc 2011 Jul 1;18(4):358–66.
6. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al. The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics 2011;4(1):13.
7. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med 2013 15(10):761-71.
8. Denny JC, Ritchie MD, Crawford DC, Schildcrout JS, Ramirez AH, Pulley JM, et al. Identification of genomic predictors of atrioventricular conduction: using electronic medical records as a tool for genome science. Circulation 2010 Nov 16;122(20):2016–21.
9. Kullo IJ, Ding K, Jouni H, Smith CY, Chute CG. A genome-wide association study of red blood cell traits using the electronic medical record. PloS One [Internet]. 2010 [cited 2010 Oct 29];5(9). Available from: http://www.ncbi.nlm.nih.gov/pubmed/20927387
10. Ritchie MD, Denny JC, Crawford DC, Ramirez AH, Weiner JB, Pulley JM, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. Am J Hum Genet 2010 Apr 9;86(4):560–72.
11. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics 2010 May 1;26(9):1205–10.
12. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inform Assoc 2013 Mar 26;
13. Denny JC, Crawford DC, Ritchie MD, Bielinski SJ, Basford MA, Bradford Y, et al. Variants Near FOXE1 Are Associated with Hypothyroidism and Other Thyroid Conditions: Using Electronic Medical Records for Genome- and Phenome-wide Studies. Am J Hum Genet 2011 Oct 7;89(4):529–42.
14. Thompson WK, Rasmussen LV, Pacheco JA, Peissig PL, Denny JC, Kho AN, et al. An evaluation of the NQF Quality Data Model for representing Electronic Health Record driven phenotyping algorithms. AMIA Annu Symp Proc 2012;2012:911–20.
15 Rea S, Pathak J, Savova G, Oniki TA, Westberg L, Beebe CE, et al. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the SHARPn project. J Biomed Inform 2012 Aug;45(4):763–71.
16. Delaney JT, Ramirez AH, Bowton E, Pulley JM, Basford MA, Schildcrout JS, et al. Predicting clopidogrel response using DNA samples linked to an electronic health record. Clin Pharmacol Ther 2012 Feb;91(2):257–63.
17. Xu H, Jiang M, Oetjens M, Bowton EA, Ramirez AH, Jeff JM, et al. Facilitating pharmacogenetic studies using electronic health records and natural-language processing: a case study of warfarin. J Am Med Inform Assoc 2011 Aug;18(4):387–91.
18. Wei W-Q, Feng Q, Jiang L, Waitara MS, Iwuchukwu OF, Roden DM, et al. Characterization of Statin Dose Response in Electronic Medical Records. Clin Pharmacol Ther [Internet] 2013 Nov 13 [cited 2013 Dec 31]; Available from: http://www.nature.com/clpt/journal/vaop/ncurrent/full/clpt2013202a.html
19. Wilke RA, Xu H, Denny JC, Roden DM, Krauss RM, McCarty CA, et al. The emerging role of electronic medical records in pharmacogenomics. Clin Pharmacol Ther 2011 Mar;89(3):379–86.
20. Bowton E, Field JR, Wang S, Schildcrout JS, Van Driest SL, Delaney JT, et al. Biobanks and electronic medical records: enabling cost-effective research. Sci Transl Med 2014 Apr 30;6(234):234cm3.
21. Hebbring SJ. The Challenges, Advantages, and Future of Phenome-Wide Association Studies. Immunology 2013 Oct 22;
22. Hebbring SJ, Schrodi SJ, Ye Z, Zhou Z, Page D, Brilliant MH. A PheWAS approach in studying HLA-DRB1*1501. Genes Immun 2013 Apr;14(3):187–91.
23. Ritchie MD, Denny JC, Zuvich RL, Crawford DC, Schildcrout JS, Bastarache L, et al. Genome- and phenome-wide analyses of cardiac conduction identifies markers of arrhythmia risk. Circulation 2013 Apr 2;127(13):1377–85.
24. Shameer K, Denny JC, Ding K, Jouni H, Crosslin DR, de Andrade M, et al. A genome- and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects. Hum Genet 2013 Sep 12;
25. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotechnol 2013 Nov 24;31(12):1102–11.
26. Pendergrass SA, Brown-Gentry K, Dudek S, Frase A, Torstenson ES, Goodloe R, et al. Phenome-wide association study (PheWAS) for detection of pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. PLoS Genet 2013;9(1):e1003087.
27. Pendergrass SA, Dudek SM, Crawford DC, Ritchie MD. Visually integrating and exploring high throughput Phenome-Wide Association Study (PheWAS) results using PheWAS-View. BioData Min 2012;5(1):5.
28. Million Veteran Program (MVP) [Internet]. [cited 2012 Jun 20]. Available from: http://www.research.va.gov/mvp/
29. The Research Program on Genes, Environment, and Health [Internet]. [cited 2014 Jan 2]. Available from: http://www.dor.kaiser.org/external/DORExternal/rpgeh/index.aspx?ekmensel=194f64c3_47_48_btnlink
30. Collins R. What makes UK Biobank special? Lancet 2012 Mar 31;379(9822):1173–4.
31. Chen Z, Chen J, Collins R, Guo Y, Peto R, Wu F, et al. China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. Int J Epidemiol 2011 Dec;40(6):1652–66.
32. LePendu P, Iyer SV, Bauer-Mehren A, Harpaz R, Mortensen JM, Podchiyska T, et al. Pharmacovigilance Using Clinical Notes. Clin Pharmacol Ther [Internet] 2013 Apr 10 [cited 2013 Apr 12]; Available from: http://www.nature.com/clpt/journal/vaop/ncurrent/full/clpt201347a.html
33. Leeper NJ, Bauer-Mehren A, Iyer SV, LePendu P, Olson C, Shah NH. Practice-Based Evidence: Profiling the Safety of Cilostazol by Text-Mining of Clinical Notes. PLoS ONE 2013 May 23;8(5):e63499.
34. Ryan PB, Madigan D, Stang PE, Schuemie MJ, Hripcsak G. Medication-Wide Association Studies. CPT Pharmacomet Syst Pharmacol 2013 Sep 18;2(9):e76.
35. Frankovich J, Longhurst CA, Sutherland SM. Evidence-Based Medicine in the EMR Era. N Engl J Med 2011;365(19):1758–9.
36. Liao KP, Kurreeman F, Li G, Duclos G, Murphy S, P RG, et al. Autoantibodies, autoimmune risk alleles and clinical associations in rheumatoid arthritis cases and non-RA controls in the electronic medical records. Arthritis Rheum 2012 Dec 10;
37. Neuraz A, Chouchana L, Malamut G, Le Beller C, Roche D, Beaune P, et al. Phenome-Wide Association Studies on a Quantitative Trait: Application to TPMT Enzyme Activity and Thiopurine Therapy in Pharmacogenomics. PLoS Comput Biol 2013 Dec 26;9(12):e1003405.
38. Boland MR, Hripcsak G, Albers DJ, Wei Y, Wilcox AB, Wei J, et al. Discovering medical conditions associated with periodontitis using linked electronic health records. J Clin Periodontol 2013 May;40(5):474–82.
39. Doshi-Velez F, Ge Y, Kohane I. Comorbidity Clusters in Autism Spectrum Disorders: An Electronic Health Record Time-Series Analysis. Pediatrics 2013 Dec 9;peds.2013–0819.
40. Blair DR, Lyttle CS, Mortensen JM, Bearden CF, Jensen AB, Khiabanian H, et al. A nondegenerate code of deleterious variants in Mendelian loci contributes to complex disease risk. Cell 2013 Sep 26;155(1):70–80.
41. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc 2013 Jan 1;20(1):117–21.
42. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. J Biomed Inform 2013 Oct;46(5):830–6.

43. Wei W-Q, Leibson CL, Ransom JE, Kho AN, Caraballo PJ, Chai HS, et al. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. J Am Med Inform Assoc 2012 Apr;19(2):219–24.
44. Collins F. Opportunities and challenges for the NIH--an interview with Francis Collins. Interview by Robert Steinbrook. N Engl J Med 2009 Oct 1;361(14):1321–3.
45. Sarkar IN. Bringing genome tests into clinical practice. Yearb Med Inform 2013;8(1):172–4.
46. Pulley JM, Denny JC, Peterson JF, Bernard GR, Vnencak-Jones CL, Ramirez AH, et al. Operational Implementation of Prospective Genotyping for Personalized Medicine: The Design of the Vanderbilt PREDICT Project. Clin Pharmacol Ther 2012 May 16;92(1):87–95.
47. Hicks JK, Crews KR, Hoffman JM, Kornegay NM, Wilkinson MR, Lorier R, et al. A Clinician-Driven Automated System for Integration of Pharmacogenetic Interpretations Into an Electronic Medical Record. Clin Pharmacol Ther 2012;92(5):563–6.
48. Van Driest SL, Shi Y, Bowton EA, Schildcrout JS, Peterson JF, Pulley J, et al. Clinically actionable genotypes among 10,000 patients with preemptive pharmacogenomic testing. Clin Pharmacol Ther 2013 Nov 19;
49. O'Donnell PH, Bush A, Spitz J, Danahey K, Saner D, Das S, et al. The 1200 patients project: creating a new medical model system for clinical implementation of pharmacogenomics. Clin Pharmacol Ther 2012 Oct;92(4):446–9.
50. Johnson JA, Elsey AR, Clare-Salzler MJ, Nessl D, Conlon M, Nelson DR. Institutional Profile: University of Florida and Shands Hospital Personalized Medicine Program: clinical implementation of pharmacogenetics. Pharmacogenomics 2013 May;14(7):723–6.
51. Gottesman O, Scott SA, Ellis SB, Overby CL, Ludtke A, Hulot J-S, et al. The CLIPMERGE PGx Program: clinical implementation of personalized medicine through electronic health records and genomics-pharmacogenomics. Clin Pharmacol Ther 2013 Aug;94(2):214–7.
52. Implementing Genomics in Practice (IGNITE) [Internet]. [cited 2014 Jan 9]. Available from: http://www.genome.gov/27554264
53. Relling MV, Klein TE. CPIC: Clinical Pharmacogenetics Implementation Consortium of the Pharmacogenomics Research Network. Clin Pharmacol Ther 2011 Mar;89(3):464–7.
54. List Dosing Guidelines [PharmGKB] [Internet]. [cited 2014 Jan 9]. Available from: http://www.pharmgkb.org/view/dosing-guidelines.do?source=CPIC
55. Scott SA, Sangkuhl K, Gardner EE, Stein CM, Hulot J-S, Johnson JA, et al. Clinical Pharmacogenetics Implementation Consortium guidelines for cytochrome P450-2C19 (CYP2C19) genotype and clopidogrel therapy. Clin Pharmacol Ther 2011 Aug;90(2):328–32.
56. Scott SA, Sangkuhl K, Stein CM, Hulot J-S, Mega JL, Roden DM, et al. Clinical Pharmacogenetics Implementation Consortium guidelines for CYP2C19 genotype and clopidogrel therapy: 2013 update. Clin Pharmacol Ther 2013 Sep;94(3):317–23.
57. Pharmacogenomic Biomarkers in Drug Labels [Internet]. [cited 2011 Mar 9]. Available from: http://www.fda.gov/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics/ucm083378.htm
58. Scannell JW, Blanckley A, Boldon H, Warrington B. Diagnosing the decline in pharmaceutical R&D efficiency. Nat Rev Drug Discov 2012 Mar;11(3):191–200.
59. Dudley JT, Sirota M, Shenoy M, Pai RK, Roedder S, Chiang AP, et al. Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. Sci Transl Med 2011 Aug 17;3(96):96ra76.
60. Sirota M, Dudley JT, Kim J, Chiang AP, Morgan AA, Sweet-Cordero A, et al. Discovery and preclinical validation of drug indications using compendia of public gene expression data. Sci Transl Med 2011 Aug 17;3(96):96ra77.
61. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science 2006 Sep 29;313(5795):1929–35.
62. Jahchan NS, Dudley JT, Mazur PK, Flores N, Yang D, Palmerton A, et al. A drug repositioning approach identifies tricyclic antidepressants as inhibitors of small cell lung cancer and other neuroendocrine tumors. Cancer Discov 2013 Dec;3(12):1364–77.
63. Sanseau P, Agarwal P, Barnes MR, Pastinen T, Richards JB, Cardon LR, et al. Use of genome-wide association studies for drug repositioning. Nat Biotechnol 2012 Apr;30(4):317–20.
64. Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature [Internet] 2013 Dec 25 [cited 2013 Dec 28];advance online publication. Available from: http://www.nature.com/nature/journal/vaop/ncurrent/full/nature12873.html
65. Khatri P, Roedder S, Kimura N, De Vusser K, Morgan AA, Gong Y, et al. A common rejection module (CRM) for acute rejection across multiple organs identifies novel therapeutics for organ transplantation. J Exp Med 2013 Oct 21;210(11):2205–21.
66. Sateriale A, Bessoff K, Sarkar IN, Huston CD. Drug repurposing: mining protozoan proteomes for targets of known bioactive compounds. J Am Med Inform Assoc 2013 Jun 11;
67. Griffith M, Griffith OL, Coffman AC, Weible JV, McMichael JF, Spies NC, et al. DGIdb: mining the druggable genome. Nat Methods 2013 Dec;10(12):1209–10.
68. Tenenbaum J, James A, Paulyson-Nuñez K. An Altered Treatment Plan Based on Direct to Consumer (DTC) Genetic Testing: Personalized Medicine from the Patient/Pin-cushion Perspective. J Pers Med 2012 Oct 30;2(4):192–200.
69. Kalf RRJ, Mihaescu R, Kundu S, de Knijff P, Green RC, Janssens ACJW. Variations in predicted risks in personal genome testing for common complex diseases. Genet Med [Internet] 2013 Jun 27 [cited 2013 Dec 27]; Available from: http://www.nature.com/gim/journal/vaop/ncurrent/full/gim201380a.html
70. Tung JY, Do CB, Hinds DA, Kiefer AK, Macpherson JM, Chowdry AB, et al. Efficient replication of over 180 genetic associations with self-reported medical data. PloS One 2011;6(8):e23473.
71. Hinds DA, McMahon G, Kiefer AK, Do CB, Eriksson N, Evans DM, et al. A genome-wide association meta-analysis of self-reported allergy identifies shared and allergy-specific susceptibility loci. Nat Genet 2013 Aug;45(8):907–11.
72. Eriksson N, Tung JY, Kiefer AK, Hinds DA, Francke U, Mountain JL, et al. Novel associations for hypothyroidism include known autoimmune risk loci. PloS One 2012;7(4):e34442.
73. 2013 - 23andMe, Inc. 11/22/13 [Internet]. [cited 2013 Dec 27]. Available from: http://www.fda.gov/ICECI/EnforcementActions/WarningLetters/2013/ucm376296.htm
74. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HYK, Chen R, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell 2012 Mar 16;148(6):1293–307.
75. Lovly CM, Dahlman KB, Fohn LE, Su Z, Dias-Santagata D, Hicks DJ, et al. Routine multiplex mutational profiling of melanomas enables enrollment in genotype-driven therapeutic trials. PloS One 2012;7(4):e35309.
76. Dahlman KB, Xia J, Hutchinson K, Ng C, Hucks D, Jia P, et al. BRAF(L597) mutations in melanoma are associated with sensitivity to MEK inhibitors. Cancer Discov 2012 Sep;2(9):791–7.
77. Tiacci E, Trifonov V, Schiavoni G, Holmes A, Kern W, Martelli MP, et al. BRAF mutations in hairy-cell leukemia. N Engl J Med 2011 Jun 16;364(24):2305–15.
78. Hairy Cell Leukemia—New Genes, New Targets - Springer. [cited 2014 Jan 4]; Available from: http://link.springer.com/article/10.1007%2Fs11899-013-0167-0/fulltext.html#Sec17
79. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. Emerging landscape of oncogenic signatures across human cancers. Nat Genet 2013 Oct;45(10):1127–33.
80. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature 2012 Oct 4;490(7418):61–70.
81. Snitkin ES, Zelazny AM, Thomas PJ, Stock F, NISC Comparative Sequencing Program Group, Henderson DK, et al. Tracking a hospital outbreak of carbapenem-resistant Klebsiella pneumoniae with whole-genome sequencing. Sci Transl Med 2012 Aug 22;4(148):148ra116.
82. Chin C-S, Sorenson J, Harris JB, Robins WP, Charles RC, Jean-Charles RR, et al. The Origin of the Haitian Cholera Outbreak Strain. N Engl J Med 2011 Jan 6;364(1):33–42.
83. Hasan NA, Choi SY, Eppinger M, Clark PW, Chen A, Alam M, et al. Genomic diversity of 2010 Haitian cholera outbreak strains. Proc Natl Acad Sci U S A 2012 Jul 17;109(29):E2010–2017.
84. Starren J, Williams MS, Bottinger EP. Crossing the omic chasm: a time for omic ancillary systems. JAMA J Am Med Assoc 2013 Mar 27;309(12):1237–8.

Correspondence to:
Joshua C. Denny, MD, MS
2525 West End Ave - Suite 672
Nashville, TN 37213, USA
E-mail: josh.denny@vanderbilt.edu