Published in final edited form as:

Nat Biotechnol. 2013 December ; 31(12): 1095–1097. doi:10.1038/nbt.2757.

Mining the ultimate phenome repository

Nigam H Shah

Nigam H. Shah is at the Center for Biomedical Informatics Research, Stanford, California, USA.
nigam@stanford.edu

Abstract

Combining genotyping and the data locked in medical records yields a large number of known genotype-phenotype associations.

The phenome-wide association study (PheWAS) is a relatively new approach for exploring the influence of genetic variation on disease etiology. First described by Denny *et al.*¹ in 2010, PheWAS involves searching for associations between genetic variants and a wide spectrum of disease phenotypes, typically captured in electronic medical records or epidemiologic studies. But how reliable are the resulting associations? In a report in this issue, Denny *et al.*² sought to validate PheWAS by comparing the associations it discovers to those identified by genome-wide association studies (GWAS). Notably, they find that a single PheWAS replicates 66% of the associations produced by multiple GWAS studies that were sufficiently powered. Moreover, PheWAS reveals new and potentially pleiotropic associations that could not have been readily identified by GWAS.

A GWAS usually starts with a phenotype chosen by the investigator and searches the genome for associated genetic variants. In contrast, PheWAS begins with genetic variants and looks for associations with a large collection of disease phenotypes (Fig. 1a). Previous PheWAS papers have studied a small number of genetic variants: two began with a genomic region of interest (e.g., HLA-DRB1*1501 and FOXE1) and searched for associated phenotypes^{3,4}, and two others began with a phenotype of interest (e.g., platelet count and rheumatoid arthritis), searched for associated genetic markers and then performed a PheWAS to identify pleiotropic associations of those genetic markers with other phenotypes^{5,6}.

Because PheWAS can detect multiple, seemingly unrelated phenotypes associated with a single genetic variant, it facilitates the identification of pleiotropic associations. And because it can make use of information buried in electronic medical records, it offers a strategy for mining data generated in routine clinical practice.

The new report by Denny *et al.*² is more ambitious than previous PheWAS. Rather than studying a handful of genetic variants, the authors analyzed associations between a large

number of variants and a large number of phenotypes—3,144 single-nucleotide polymorphisms (SNPs) and 1,358 phenotypes extracted from electronic medical records—thereby allowing associations to be discovered in an unbiased manner (Fig. 1b). Moreover, they systematically compared the associations produced by the PheWAS with those in the National Human Genome Research Institute (NHGRI)'s GWAS Catalog, a repository for GWAS data.

The population studied by Denny *et al.*² consisted of 13,835 individuals of European descent whose samples and associated electronic medical records were available through the Electronic Medical Records and Genomic (eMERGE) Network. Funded by NHGRI, eMERGE is a national consortium designed to create, share and apply approaches to integrate DNA biorepositories with electronic medical record systems, with the ultimate goal of extracting genomic testing results and providing them to patients in a clinical care setting. To enable comparison with GWAS data, the authors restricted their analysis to SNPs in the NHGRI GWAS Catalog. At the beginning of their study, the catalog contained 6,092 SNPs. Of these, 3,144 were present and passed quality control on the Illumina chip that the authors used for genotyping. After applying a commonly accepted cut-off for genome-wide significance of $P = 5 \times 10^{-8}$, the authors narrowed down the 3,144 SNPs to 673—or 751 SNP-phenotype associations—which they went on to compare with the results of PheWAS.

Electronic medical records are generated as a by-product of health care rather than for research as the primary goal. Data are collected when a patient is ill, are split across visits and facilities, are not standardized in form, and are collected with the expectation that a human being (such as another clinician), rather than an automated system, will read the information in the record. Therefore, it can be difficult to extract phenotypic information from electronic medical record data.

To partially address these challenges, Denny *et al.*² created shareable and potentially reusable phenotype descriptions primarily comprising codes from the International Classification of Disease, ninth revision, Clinical Modification (ICD9). ICD9 codes are alphanumeric codes corresponding to the diagnoses and procedures recorded in conjunction with a doctor visit and are used primarily for billing. To expand the list of PheWAS phenotypes beyond those used in their earlier work¹, and in an effort to extract fine-grained NHGRI GWAS phenotypes from electronic medical record data, Denny *et al.*² included additional types of codes and created a hierarchy of phenotypes (e.g., introducing inflammatory bowel disease as a parent phenotype for Crohn's disease and ulcerative colitis). Although other items from an electronic medical record could potentially be used (e.g., text mentions, laboratory results and medical orders), the authors chose to define the phenotypes as ICD9 codes because these codes are the most universally available form of electronic medical record data. The phenotype definitions created by Denny *et al.*² should be useful to others seeking to conduct PheWAS or to refine phenotype descriptions by including additional data items from the electronic medical record.

To test each SNP-phenotype association, the authors identified the number of patients who had the phenotype (defined as having two distinct instances of the relevant ICD9 code) and the suitable controls (those that did not have the ICD9 codes corresponding to the

phenotype). Patients with only one occurrence of the ICD9 code were excluded from the analysis. The strength of the association between the SNP and phenotype was calculated with the PLINK tool using logistic regression adjusted for age, gender and site, assuming an additive genetic model.

Denny *et al.*² found that PheWAS replicated 210 of the 751 SNP-phenotype associations from the NHGRI GWAS Catalog. To determine why not all associations were replicated, they filtered the 751 associations according to three criteria: the statistical power of the original study, the number of independent studies that reported the association, and whether the GWAS trait exactly matched the electronic medical record phenotype used for PheWAS. The likelihood of replication increased proportionally with the statistical power of the original GWAS study, the number of independent GWAS studies reporting the association, and the degree of match between GWAS and electronic medical record traits. After adjusting for these factors, they concluded that PheWAS replicated 51 of 77 associations (66%).

Notably, PheWAS also identified 63 electronic medical record trait–genotype associations that are not included in the NHGRI GWAS Catalog and are potentially novel. Further study replicated two of these new associations in an independent patient cohort. In fact, during the validation process, the authors also expanded their phenotype definitions to include text mentions in the electronic medical record identified by natural language processing.

The main advantage of PheWAS—the ability to search for associations between SNPs and a large number of phenotypes—also results in its main limitation, namely, the potential for high false-positive rates. Denny *et al.*² controlled for such false discovery by adjusting the *P*-value cutoff to allow for 10% false discovery rate. Adjusting the *P*-value for an acceptable false discovery rate is better than a simple Bonferroni correction, which would divide the traditional *P*-value cutoff of 0.05 by the number of associations tested. Naturally, a statistical correction is no guarantee that the false discovery rate is truly 10%. It may be possible to create a ground-truth data set with known negative associations, which in conjunction with the known positive associations from GWAS could be used to quantify the actual false discovery rate for a high-throughput PheWAS study.

Electronic medical records are increasing at a rapid pace and represent the ultimate repository of disease-phenotype information. According to the Office of the National Coordinator for Health Information Technology, hospital adoption of electronic medical record systems has grown from 12% to 44% since 2009. Electronic medical record adoption by family physicians is estimated to exceed 80% by the end of 2013.

One important application of electronic medical record data is for genetics research, as in the PheWAS of Denny *et al.*². In another notable instance, Blair *et al.*⁷ used the medical records of >110 million patients to find reproducible patterns linking complex disorders to unique collections of Mendelian loci. But the utility of electronic medical records extends well beyond genetics. For example, phenotyping for clinical research, particularly drug safety research and patient-centered outcomes research, is already underway. The FDA's Sentinel initiative, launched in 2008 to track the safety of drugs, biologics and medical devices using

electronic health record systems, administrative and claims databases, and registries, already analyzes data from over 100 million patients (<http://www.mini-sentinel.org/>). The recently launched National Patient- Centered Clinical Research Network (<http://www.pcori.org/funding-opportunities/improving-our-national-infrastructure-to-conduct-comparative-effectiveness-research>) aims to build national capacity to conduct comparative effectiveness research using data from electronic medical records.

Eventually, electronic medical record data might help to bridge the ‘evidence gap’—the gap between the evidence needed to support care decisions and the evidence produced by randomized controlled trials— which was highlighted in a recent report by the Institute of Medicine. It may be possible to ‘learn’ physician practice patterns from the record of routine clinical practice embedded in electronic medical records, thereby generating practice-based evidence⁸.

More generally, electronic medical record–based phenotyping has the potential to provide a link between studies that advance the science of medicine (e.g., PheWAS) and studies that advance the practice of medicine⁹. But achieving this will require reliable methods of ‘electronic phenotyping’ that allow unambiguous recognition of disease conditions or traits from electronic medical record data. This would mean mining not only ICD9 codes but also mentions in textual notes, expected laboratory test values and drug prescriptions. As part of the eMERGE Network, Denny and others have already begun a collaboration to create the Phenome Knowledge Base (PheKB; <http://www.phekb.org/>). This resource provides an anchor for a community interested in building and validating methods for electronic medical record–based phenotyping¹⁰. PheKB already lists phenotyping methods that go beyond billing codes and text-mentions, and presents algorithms able to extract 21 different phenotypes and/or diseases from electronic medical record data.

Robust descriptions of phenotypes (referred to as ‘health outcomes of interest’ in some research communities) would allow full exploitation of data already collected in electronic medical records to advance the genomic understanding of disease, to improve clinical practice^{8,9} and to make discoveries through clinical data mining. The work of Denny *et al.*² to generate reusable, accurate and unambiguous descriptions of health outcomes or phenotypes is an important step toward this goal.

References

1. Denny JC, et al. *Bioinformatics*. 2010; 26:1205–1210. [PubMed: 20335276]
2. Denny JC, et al. *Nat. Biotechnol.* 2013; 31:1102–1104. [PubMed: 24270849]
3. Denny JC, et al. *Am. J. Hum. Genet.* 2011; 89:529–542. [PubMed: 21981779]
4. Hebbiring SJ, et al. *Genes Immun.* 2013; 14:187–191. [PubMed: 23392276]
5. Liao KP, et al. *Arthritis Rheum.* 2013; 65:571–581. [PubMed: 23233247]
6. Shameer K, et al. *Hum. Genet.* 2013
7. Blair DR, et al. *Cell.* 2013; 155:70–80. [PubMed: 24074861]
8. Leeper NJ, et al. *PLoS ONE.* 2013; 8:e63499. [PubMed: 23717437]
9. Frankovich J, Longhurst CA, Sutherland SM. *N. Engl. J. Med.* 2011; 365:1758–1759. [PubMed: 22047518]
10. Newton KM, et al. *J. Am. Med. Inform. Assoc.* 2013; 20:e1, e147–e154. [PubMed: 23687172]

Figure 1.

Approaches for deriving associations between genetic variants and disease phenotypes. Rows represent individual genetic variants (usually SNPs); columns represent individual phenotypes. **(a)** A GWAS tests one phenotype for association with many genetic variants, whereas previous PheWAS have tested associations of a genetic variant with many phenotypes. **(b)** In an expanded PheWAS, Denny *et al.*² test many genetic variants against many phenotypes extracted from electronic medical records to validate PheWAS and to

make new discoveries. Note that not all phenotypes can be reliably extracted from electronic medical record data.