A simulation study investigating power estimates in Phenome-Wide Association Studies

Anurag Verma[1,2], Yuki Bradford[1], Scott Dudek[1], Shefali S. Verma[1,2], Sarah A. Pendergrass[1], Marylyn D. Ritchie[1.2]

[1]*Biomedical and Translational Informatics, Geisinger Health System, Danville, PA*

[2]*The Huck Institutes of the Life Science, Pennsylvania State University, University Park, PA*

Abstract

Background

Phenome-Wide Association Studies (PheWAS) are a high throughput approach to evaluate comprehensive associations between genetic variants and a wide range of phenotypic measures. One of the main challenges with PheWAS is the varying sample size ranges of cases and controls across the many phenotypes of interest, that could affect statistical power to detect associations. The motivation of this study is to investigate the parameters, including sample size, that affect estimation of statistical power in PheWAS.

Results

We performed a PheWAS simulation study, where we investigated variation in statistical power based on different parameters like overall sample size, number of cases, case-control ratio, minor allele frequency, and disease penetrance. The simulation was performed on both dichotomous and continuous phenotypic measures. Our simulation on dichotomous traits suggests that the number of cases have more impact than the case to control ratio; also we find that a sample size of 200 cases or more seems to maintain statistical power to identify associations for common variants. For continuous measures, a sample size of 1000 or more individuals performed best in the power calculations. We primarily focused on common variants (MAF>0.01) in this study; However, in future studies, we will be extending this effort to perform similar simulations on rare variants.

Conclusions

This study provides a series of PheWAS simulation analysis that can be used to estimate statistical power under a number of potential scenarios. These results can be used to provide guidelines for appropriate study design for future PheWAS analyses.

Keywords

PheWAS – EHR – ICD-9 codes – Power analysis – Geisinger Health System

Background

PheWAS approaches have been implemented in variety of different studies like the eMERGE network[1–5] with electronic health record information that includes international classification of disease version 9 (ICD-9) code based diagnoses, laboratory test measurements and demographic information. Other PheWAS include data from epidemiological studies[6,7], as well as clinical trials[8,9] such as the AIDS clinical trial group (ACTG) which collected a range of measurements for different clinical domains like pharmacology, metabolism, virology, and immunology[8,9]. Cohorts like these with large number of measurements for every individual have made PheWAS an effective approach to scan over hundreds and thousands of associations in a high-throughput way. PheWAS serve as a great tool to generate genetic association hypotheses for multiple phenotypes as well as provide insights in cross-phenotype associations. Unlike a GWAS where one phenotype is evaluated, PheWAS utilizes a wide range of phenotypes collected for a variety of biological interests for each dataset. Thus in PheWAS, the data collected for different measurements can vary dramatically in sample size, including specifically the numbers of cases for diagnoses can be considerably different depending on the rarity of the diagnosis. This makes the estimation of statistical power for PheWAS a challenge. For example, in electronic health record (EHR) data, one of the most commonly used data types are ICD-9 codes; these codes provide information on disease diagnosis, procedures, and medications in the form of 3-5 digit codes. The longitudinal ICD-9 data collected over many years varies drastically between patients due to multiple factors such as differences in the frequency of patient visits, differences in length of records due to varying start and end dates, and lack of medical history with outpatient data. These factors generate sparseness and missing information in the data and hence variability in number of cases, case-control ratio, as well as overall sample size in case-control study designs. These factors could potentially affect the estimations from association testing.

The goal of this study is to perform power estimations altering number of parameters encountered in PheWAS, to determine the appropriate thresholds to use for number of cases or total sample size, as well as the ratio of cases to controls for a given phenotype. It is challenging to make study design decisions regarding sample size at the outset of a PheWAS from an EHR with either case-control phenotypes (ICD-9 codes) or quantitative trait phenotypes (clinical lab variables) as the number of cases or samples will vary for each phenotype. We run into three issues with low sample size phenotypes: 1) we have low statistical power to identify or replicate any associations and 2) we may potentially have biased estimates in analyses with low sample size, and 3) we increase our multiple hypothesis testing burden by testing low powered phenotypes. In this paper, we investigate a number of different factors that could influence the statistical power in association testing with a PheWAS and these simulations can provide empirically derived evidence to guide future PheWAS study designs.

Methods:

Simulation Study

We designed a simulation approach with different combinations of genotype and phenotype parameters and then performed association testing to investigate the factors that could influence the statistical power to detect a signal.

The design process for a binary disease outcome was to simulate datasets with different number of cases, and for each count of cases we evaluated all permutations of parameters including: case to control ratio, minor allele frequency, and several disease penetrance measures. We used R to generate random population-based samples with genotypes and their disease status using different input parameters aforementioned. For example, we generated one simulated model with 10 cases and 10 controls when case to control ratio is 1:1, 0.01 minor allele frequency, 0.15 disease penetrance and an additive genotype model (0,1,2). Under each simulation model, we generated 1000 datasets for each combination and then calculated associations using logistic regression for binary traits and linear regression for quantitative traits implemented in the statistical package PLATO (http://ritchielab.psu.edu/plato). Please refer to **Table 1** for all the different combinations of parameter values used for simulation.

For the continuous or quantitative trait simulations, we investigated the power estimates by varying the sample size, and for each bin we evaluated the power for different permutations of minor allele frequency and disease penetrance. Again, we generated 1000 datasets for each combination and then used linear regression to calculate associations with the quantitative traits. Please refer to **Table 2** for all the different combinations of parameter values used for the quantitative trait simulations.

Table 1: Parameters for case-control simulation

Case-Control Ratio	Cases	Controls	MAF	Penetrance	β_0
1:2 1:4 1:8 1:16 1:50 1:100 1:500 1:1000	10	20 40 80 160 500 1,000 5,000 10,000	0.01 0.05 0.1 0.25 0.4	0.15 0.2 0.3	0.1
1:2 1:4 1:8 1:16 1:50 1:100 1:500 1:1000	50	100 200 400 800 2,500 5,000 25,000 50,000	0.01 0.05 0.1 0.25 0.4	0.15 0.2 0.3	0.1
1:2 1:4 1:8 1:16 1:50 1:100 1:500 1:1000	100	200 400 800 1,600 5,000 10,000 50,000 100,000	0.01 0.05 0.1 0.25 0.4	0.15 0.2 0.3	0.1
1:2 1:4 1:8 1:16 1:50 1:100 1:500 1:1000	200	400 800 1,600 3,200 10,000 20,000 100,000 200,000	0.01 0.05 0.1 0.25 0.4	0.15 0.2 0.3	0.1
1:2 1:4 1:8 1:16 1:50	500	1,000 2,000 4,000 8,000 25,000	0.01 0.05 0.1 0.25 0.4	0.15 0.2 0.3	0.1

1:100 1:500 1:1000		50,000 250,000 500,000			
1:2 1:4 1:8 1:16 1:50 1:100 1:500 1:1000	1000	2,000 4,000 8,000 16,000 50,000 100,000 500,000 1,000,000	0.01 0.05 0.1 0.25 0.4	0.15 0.2 0.3	0.1

Table 2: Parameters for quantitative measurements

Sample Size	MAF	Penetrance	β_0
10	0.01, 0.05, 0.1, 0.25, 0.4	0.15, 0.2, 0.3	0.1
50	0.01, 0.05, 0.1, 0.25, 0.4	0.15, 0.2, 0.3	0.1
100	0.01, 0.05, 0.1, 0.25, 0.4	0.15, 0.2, 0.3	0.1
200	0.01, 0.05, 0.1, 0.25, 0.4	0.15, 0.2, 0.3	0.1
500	0.01, 0.05, 0.1, 0.25, 0.4	0.15, 0.2, 0.3	0.1
1,000	0.01, 0.05, 0.1, 0.25, 0.4	0.15, 0.2, 0.3	0.1
5,000	0.01, 0.05, 0.1, 0.25, 0.4	0.15, 0.2, 0.3	0.1
10,000	0.01, 0.05, 0.1, 0.25, 0.4	0.15, 0.2, 0.3	0.1
25,000	0.01, 0.05, 0.1, 0.25, 0.4	0.15, 0.2, 0.3	0.1

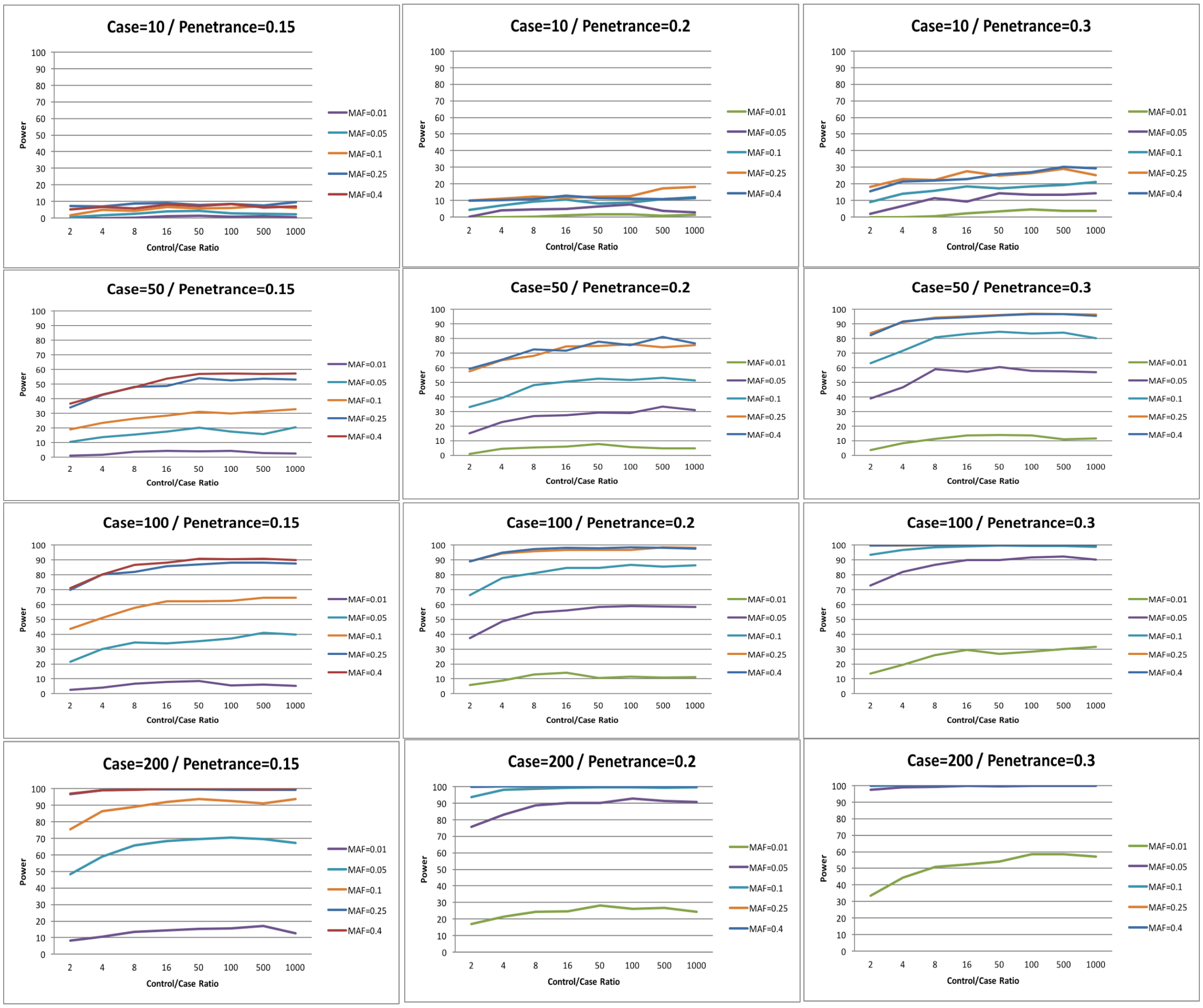
Results

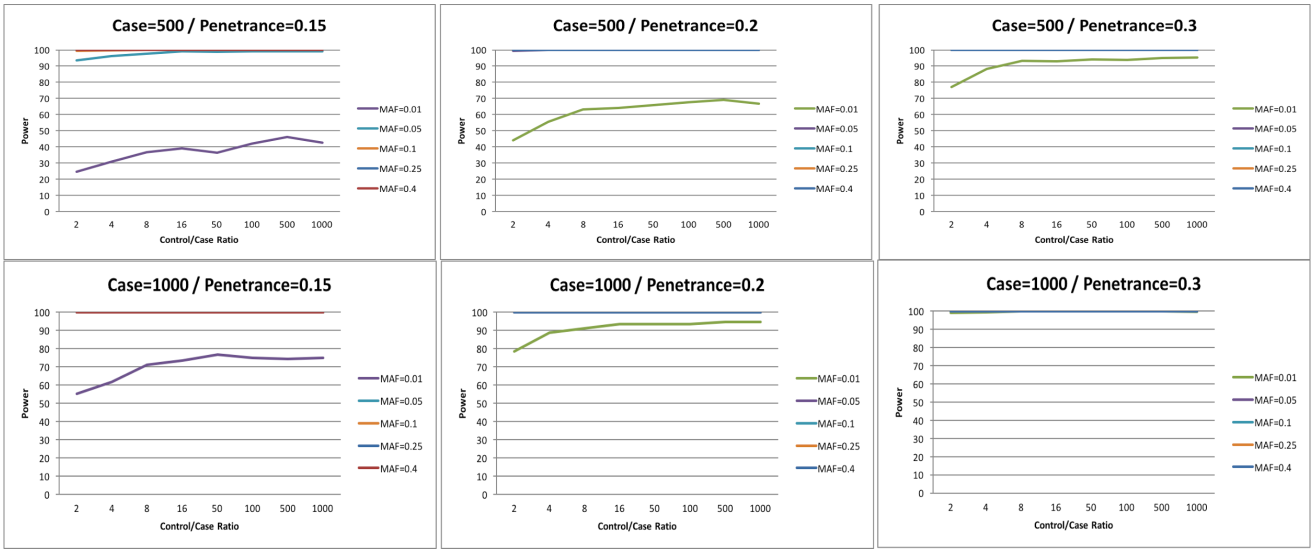
Binary variable simulation

We designed a simulation approach with different combinations of genotype and phenotype parameters and then performed association testing to investigate the factors that could influence power to detect a signal as explained ablove. We performed all permutations of the parameter settings shown in **Table 1**.

In **Figure 1**, we show trends in the estimates of power at a p-value of 0.01 for different parameters used for simulation. First, we observed an increase in power with an increase in penetrance irrespective of any change in other parameters and this is expected as highly penetrant disease are more likely to be identified even with small numbers of samples (this is due to having a high effect size). We also identified that the ratio between cases and controls does not have much impact on the power. It is actually the number of cases that largely influences the power to detect genetic associations. For example, as shown in **Figure 1**, the case-control ratio has inconsiderable affect on power whereas with the increase in case number we see three times increase in power to identify an association. These simulations also show the importance of minor allele frequency threshold when calculating associations on genotype models with an additive effect. Here, we find that all of the simulation models showed increased performance with minor allele frequency greater than 5%. The model with lower frequency variants (MAF between 1% and 5%) did not reach 100% power until the case threshold was increased to 1000 samples and the model exhibited high disease penetrance.

Figure 1. Binary Trait Power Results. Power of each simulation analysis with case-control ratio on the x-axis, minor allele frequency indicated by different colored lines, and power on the y-axis. Each box presents power for one value of the case count and penetrance combination.

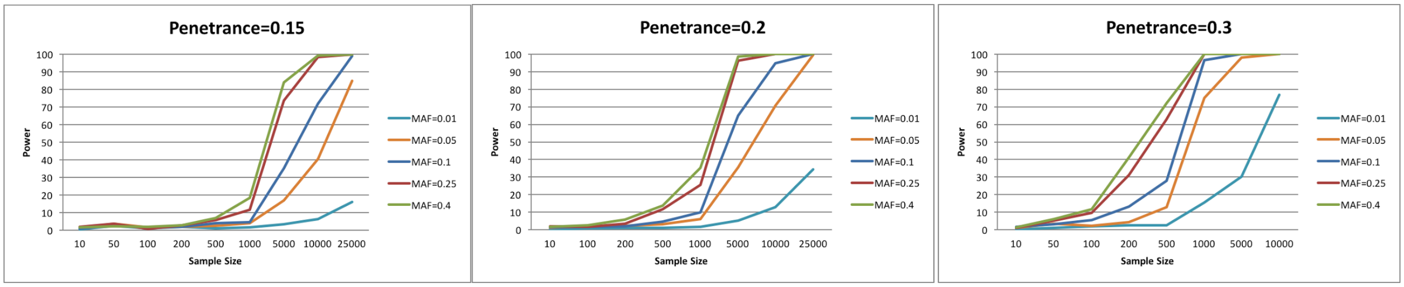




Quantitative trait simulation

We also performed similar simulation analysis on quantitative measurements (such as clinical lab variables) to identify a sample size threshold for multi-phenotype based studies like PheWAS. We used the same parameter matrix for simulations as was done for binary variables with the exception of the cases and case-control ratio (**Table 2**). For quantitative traits, we use different sample size numbers for the simulation ranging from 10 to 25,000, as these are based on estimates of sample sizes we observe in EHR or clinical trials datasets. In **Figure 2**, we show a similar graph as in the binary variable simulation where the x-axis has different sample sizes of the data set and on y-axis is the power of association. We observed almost no power until the dataset had approximately 1000 samples for a phenotype with penetrance of 0.15 and as expected we see the increase in power with higher penetrance even at smaller sample sizes. Around the sample size of 1000, we see an increase in power with slight variation with different minor allele frequencies. Again variants with rare minor alleles did not perform well until a sample size of 1000 and penetrance of 0.3. These quantitative variable simulations suggest that a threshold of 1000 samples for models with MAF greater than 5% in PheWAS and larger sample size or different statistical approach to evaluate rare variants.

Figure 2: Quantitative Trait Power Results. power of each simulation analysis by sample size, penetrance and minor allele frequency.



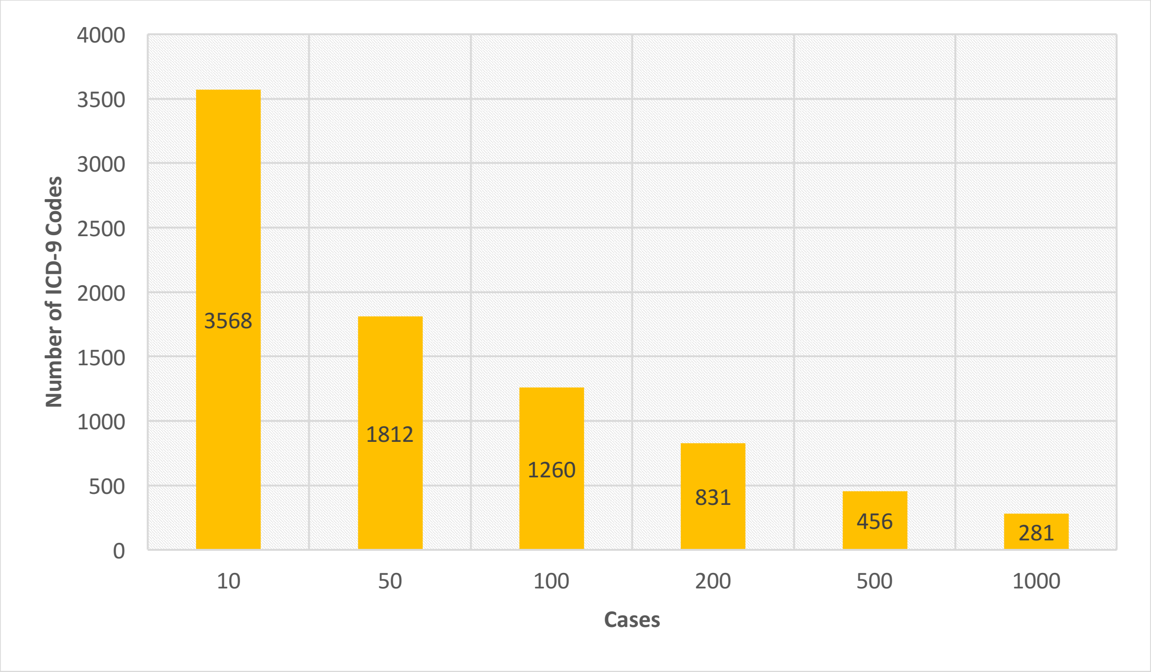
Discussion

PheWAS provides a great genomic landscape of multiple phenotypes, but a challenge of PheWAS is the range of sample sizes and case numbers that is inherent with large EHR or clinical trials-based datasets. For example, there are 14,025 possible ICD-9 diagnosis codes and 3,824 procedure codes used by EHR systems within healthcare provider organizations. With the introduction of ICD-10, the number of ICD-based codes has further increased to approximately 69,000. PheWAS is a high throughput approach to test for hundreds and thousands of variables, but testing 14,025 diagnosis codes for association with up to 1-million or more genetic variants can result in a very high multiple testing burden. Also, a large fraction of codes will have very low case numbers due to rarity of the disease, and thus may not be sufficiently powered for association detection. For example, Geisinger Heath System (GHS) is one of the largest health care providers in central Pennsylvania with an EHR including ~1.2 million unique patients. We looked into the GHS EHR data of around 100,000 participants consented into the MyCode® Community Health Initiative[10] to evaluate the extent of the variability in number of ICD-9 codes by case sample count. In order to account for misdiagnosis of missing medical history, it is advisable to define a patient as a case for an ICD-9 code only when they have three or more visits in their EHR where that specific code was represented in the patient record. Out of 14,025 codes, 33% are not coded at all and ~30% fewer than 10 patients with that code (case count <10). As shown in **Figure 3**, even after dropping out more than 60% of the ICD-9 codes there are still 3,568 codes with 10 or more patients labeled as cases, which still adds to the multiple hypothesis burden. As we identified in our binary trait simulation, the power for identifying association shows considerable increase at a case threshold of 200 with power estimates > 50% for common variant (MAF > 0.01). In the GHS EHR data, there are 831 ICD-9 codes with at least 200 patients labelled as case (as shown in **Figure 3**). We recommend using 200 as a case threshold for a common variant PheWAS analysis as it provides enough power to identify the association and also reduces multiple hypothesis burden by excluding low confidence ICD-9 codes. In the case of PheWAS on quantitative traits, our simulation suggests that a sample size of 1000 individuals will provide enough power to identify an association.

Conclusions

PheWAS have become a common tool to explore the genotype-phenotype landscape of large biobanks linked to comprehensive phenotype/trait data collections as in EHRs, clinical trials, or epidemiological cohort studies. This high-throughput analysis approach has been met with much success in recent years (cite some PheWAS papers from the literature). However, the community has been lacking guidance for making study design decisions regarding sample size, case to control ratios, and minor allele frequency. At present, there is not a PheWAS Power Calculator available to researchers. Thus, we implemented a large-scale simulation study to provide some guidelines for understanding the statistical power of PheWAS analyses under different scenarios. We believe these simulation results provide the needed power estimates for future PheWAS analysis decisions.

Figure 3 Frequency distribution of ICD-9 code-based case counts in 100,000 MyCode participants EHRs



Acknowledgments and Declarations

We would like to thank Dr. Tooraj Mirshahi and Dr. Janet Robishaw for helpful discussion on this manuscript.This work was funded in part by the following: AI077505, GM111913, HG008679, and SAP 4100070267. This project is funded, in part, under a grant with the Pennsylvania Department of Health. The Department specifically disclaims responsibility for any analyses, interpretations or conclusions.

References

1. 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. doi:10.1186/1755-8794-4-13

2. 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 Off J Am Coll Med Genet. 2013;15: 761–771. doi:10.1038/gim.2013.72

3. Ritchie MD, Verma SS, Hall MA, Goodloe RJ, Berg RL, Carrell DS, et al. Electronic medical records and genomics (eMERGE) network exploration in cataract: Several new potential susceptibility loci [Internet]. 19 Sep 2014. Available: http://www.molvis.org/molvis/v20/1281/

4. Verma A, Verma SS, Pendergrass SA, Crawford DC, Crosslin DR, Kuivaniemi H, et al. eMERGE Phenome-Wide Association Study (PheWAS) identifies clinical associations and pleiotropy for stop-gain variants. BMC Med Genomics. 2016;9. doi:10.1186/s12920-016-0191-8

5. Verma A, Basile AO, Bradford Y, Kuivaniemi H, Tromp G, Carey D, et al. Phenome-Wide Association Study to Explore Relationships between Immune System Related Genetic Loci and Complex Traits and Diseases. Yao Y-G, editor. PLOS ONE. 2016;11: e0160573. doi:10.1371/journal.pone.0160573

6. 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. doi:10.1371/journal.pgen.1003087

7. Hall MA, Verma A, Brown-Gentry KD, Goodloe R, Boston J, Wilson S, et al. Detection of Pleiotropy through a Phenome-Wide Association Study (PheWAS) of Epidemiologic Data as Part of the Environmental Architecture for Genes Linked to Environment (EAGLE) Study. Gibson G, editor. PLoS Genet. 2014;10: e1004678. doi:10.1371/journal.pgen.1004678

8. Verma A, Bradford Y, Verma SS, Pendergrass SA, Daar ES, Venuto C, et al. Multiphenotype association study of patients randomized to initiate antiretroviral regimens in AIDS Clinical Trials Group protocol A5202: Pharmacogenet Genomics. 2017;27: 101–111. doi:10.1097/FPC.0000000000000263

9. Moore CB, Verma A, Pendergrass S, Verma SS, Johnson DH, Daar ES, et al. Phenome-wide Association Study Relating Pretreatment Laboratory Parameters With Human Genetic Variants in AIDS Clinical Trials Group Protocols. Open Forum Infect Dis. 2015;2: ofu113-ofu113. doi:10.1093/ofid/ofu113

10. Carey DJ, Fetterolf SN, Davis FD, Faucett WA, Kirchner HL, Mirshahi U, et al. The Geisinger MyCode community health initiative: an electronic health record–linked biobank for precision medicine research. Genet Med. 2016;18: 906–913. doi:10.1038/gim.2015.187