Methodology in phenome-wide association studies

Citation for published version:
Wang, L, Zhang, X, Meng, X, Koskeridis, F, Georgiou, A, Yu, L, Campbell, H, Theodoratou, E & Li, X 2021, 'Methodology in phenome-wide association studies: a systematic review', *Journal of Medical Genetics*, vol. 58, no. 11. https://doi.org/10.1136/jmedgenet-2021-107696

Digital Object Identifier (DOI):
10.1136/jmedgenet-2021-107696

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Journal of Medical Genetics

Title: A systematic review of the methodology in phenome wide association studies

Authors: Lijuan Wang[1], Xiaomeng Zhang[2], Xiangrui Meng[3], Fotios Koskeridis[4], Andrea Georgiou[4], Lili Yu[1], Harry Campbell[2], Evropi Theodoratou[2, 5*], Xue Li[1, 2 *]

[1]School of Public Health and the Second affiliated hospital, Zhejiang University, Hangzhou, China.
[2]Centre for Global Health, Usher Institute, University of Edinburgh, Edinburgh, UK.
[3]Vanke School of Public Health, Tsinghua University, Beijing, China.
[4]Department of Hygiene and Epidemiology, University of Ioannina Medical School, 45110, Ioannina, Greece.
[5]Cancer Research UK Edinburgh Centre, Medical Research Council Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK.

Word count: 4930.

[*] The last two authors should be regarded as Joint Last Authors.

Correspondence to:

Xue Li, School of Public Health and the Second Affiliated Hospital, Zhejiang University, Hangzhou, China, Email: xueli157@zju.edu.cn, Tel: +86 181 5714 0559

Evropi Theodoratou, Centre for Global Health, Usher Institute, The University of Edinburgh, United Kingdom, Email: e.theodoratou@ed.ac.uk, Tel: + 44 0131 650 3210

Abstract

Background: Phenome wide association study (PheWAS) has been increasingly used to identify novel genetic associations across a wide spectrum of phenotypes. This systematic review aims to summarize the PheWAS methodology, discuss the advantages and challenges of PheWAS and provide potential implications for future PheWAS studies.

Methods: MEDLINE and EMBASE databases were searched to identify all published PheWAS studies up until April 24, 2021. A summary of PheWAS methodology incorporating how to perform PheWAS analysis and which software/tool could be used, were summarized based on the extracted information.

Results: A total of 1,035 studies were identified and 195 eligible articles were finally included. Among them, 137 (77.0%) contained 10,000 or more study participants, 164 (92.1%) defined the phenome based on electronic medical records (EMR) data, 140 (78.7%) used genetic variants as predictors, and 73 (41.0%) conducted replication analysis to validate PheWAS findings and almost all of them (94.5%) received consistent results. The methodology applied in these PheWAS studies was dissected into several critical steps, including quality control of the phenome, selecting predictors, phenotyping, statistical analysis, interpretation and visualization of PheWAS results, and the workflow for performing a PheWAS was established with detailed instructions on each step.

Conclusions: This study provides a comprehensive overview of PheWAS methodology to help practitioners achieve a better understanding of the PheWAS design, to detect understudied or overstudied outcomes and to direct their research by applying the most appropriate software and online tools for their study data structure.

Keywords: phenome-wide association study, electronic medical record, pleiotropy, genome-wide association study

Introduction

Genome wide association study (GWAS) has been a standard method for exploring the genetic etiology of common complex diseases. [1] As of April 24, 2021, GWASs have identified 257,351 genetic associations from 5,037 publications based on the GWAS catalog database.[2] Further post-GWAS analyses, when integrating biological repositories and clinical information, provide important insights in discovering novel biological mechanisms and clinical applications.[3] Generally, although GWAS makes great contributions to explore the genetic profiles for complex traits, it restricts the research to one or a small set of diseases.

To complement GWAS, the concept of phenome wide association study (PheWAS), a reversal of GWAS paradigm, has been proposed. PheWAS aims to identify the associations between an exposure (genetic variants, genetic risk score, biomarkers or risk factors) with a broad range of human phenotypes.[4] Due to the wide availability of dense electronic medical records (EMR), PheWAS analysis has started to be widely adopted by using EMR as an efficient data source for phenotype extraction.[5] Furthermore, a number of large biobanks such as the UK Biobank,[6] China Kadoorie Biobank,[7] the Electronic Medical Records and Genomics Network (the eMERGE Network)[8] and the Veteran Administration's Million Veteran Program (MVP),[9] have linked large volume of genotypic data to EMR data, making it possible to perform powerful PheWAS.

In 2010, Denny JC *et al*. illustrated the first application of PheWAS methodology with EMR data and successfully identified significant genotype-phenotype associations.[4] Subsequent PheWAS studies emerged to explore phenotypic associations in larger population with different data structure, leading to a rapid development of this approach. Bush WS *et al*. conducted a systematic review in 2016 to comprehensively introduce the rationale, methodology, findings and challenges of PheWAS.[10] Since then, multiple novel strategies as well as diverse software have been developed to improve the PheWAS methodology, therefore a detailed pipeline is needed to offer suggestions for the best practices of PheWAS analysis using appropriate methods. Here, we conducted a systematic review to identify all published PheWAS studies, with the aim of exploring common PheWAS design, suggesting the workflow for conducting and reporting PheWAS, and discussing the future development and application of this approach in clinical research.

Methods

Search strategy and eligible criteria

A systematic literature search was performed in MEDLINE and EMBASE database (Ovid) from inception to 24th April 2021, using a comprehensive search strategy. The search terms comprised medical subject headings (MeSH) and keywords relating to PheWAS (i.e., "phenome wide association stud*" OR "PheWAS" OR "phenome wide mendelian randomi*"). All identified publications went through a parallel review of the title, abstract and full text (performed by L.W. and F.K. independently) based on predefined inclusion and exclusion criteria. In particular, we included studies that employed PheWAS analysis to explore the relationship between exposures (genetic variants, environmental exposures and laboratory variables) and a wide range of phenotypes; and studies that introduced a new software/tool for PheWAS analysis. Conversely, we excluded (i) studies not applying PheWAS analysis; (ii) studies not aiming at exploring associations of exposures with phenotypic outcomes; and (iii) reviews, correspondence, conference abstracts, comments, survey, and research experiments conducted in animals or animal/human cell lines. When inconsistency in decision occurred, the two authors (L.W. and F.K.) discussed with the assistance of a third author (A.G.) to arrive at a consensus.

Data extraction

From each study, we extracted the following information: cohort name, sample size, ethnicity, type of data (i.e. EMR cohorts, epidemiological cohorts or clinical trials), type of predictors (genetic instruments or non-genetic risk factors), phenotyping method (i.e. ICD curated/holistic, number of PheWAS groups, case definition), multiple testing correction method, adjustment for covariates, key findings, and any other methodological improvement emphasized in their main text. Data extraction was performed by two investigators (L.W. and L.Y.) and double checked by another two investigators (F.K. and A.G.).

Data synthesis

Based on the extracted information, we presented the basic characteristics of published studies and made a comprehensive summary of the main steps of PheWAS analysis accordingly. We then conducted a narrative synthesis of the main packages and statistical software used by each step. Characteristics of these software were described and a comparison was carried out to determine the most appropriate option for PheWAS analysis with different data types.

Results

Included studies and characteristics

A total of 1,035 articles were retrieved from the systematic search in two databases. Eventually, 195 papers were eligible for inclusion (**Figure 1**). Of them, 178 articles reported original PheWAS studies and 17 articles introduced online resources and tools applicable for PheWAS analysis. Detailed information and the main findings for each study are presented in the **Supplementary Table 1** and **Supplementary Table 2**, respectively.

The main characteristics regarding the PheWAS methodology of these included papers are summarized in **Table 1**. Among them, 137 (77.0%) had very large sample size with >10,000 participants; and most of the studies (92.1%) defined the phenome based on the widely available EMR data in health care systems while the rest (7.9%) applied phenome definitions from epidemiological studies. 21 (11.8%) studies used a single nucleotide polymorphism (SNP) as predictor, and 119 (66.9%) studies selected multiple genetic variants as instruments, among which 31 (17.4%) studies constructed polygenic risk scores (PRS) with multiple SNPs and 88 (49.4%) studies used multiple variants as independent predictors. About half of the studies (41.0%) conducted replication analysis using an independent dataset or applying additional analysis such as mendelian randomization (MR) to further validate the PheWAS findings and 94.5% of them obtained consistent conclusions.

Critical steps of conducting a PheWAS

The main steps of PheWAS are summarized based on the methods applied in these eligible studies, and are shown in **Figure 2**.

Step 1: Quality control of the phenome

The first critical step of conducting a PheWAS is to do quality control (QC) of the phenome. Several issues in terms of QC have been proposed in the published studies, such as incompatibility with rare variant analysis or outcomes due to relatively small sample size, great multiple test burden resulting from a large number of phenotypes and low-quality phenotyping caused by data missingness, leading to low statistical power. Thus, a cut-off threshold used to select appropriate phenotypes or power calculations before PheWAS should be considered to maintain the statistical power. Based on the extracted information, a large proportion of PheWAS studies only included phenotypes with at least 20 cases. A simulation study investigated the effects of various parameters on the estimation of statistical power in PheWAS, and concluded that a number of 200 cases or more maintains the statistical power to identify associations for common traits.[11] Beyond this commonly used cut off threshold, there are examples of PheWAS that used specific software to maintain the statistical power. Namjou B *et al.* used QUANTO software to calculate the power for the included phenotypes before PheWAS analysis.[12] Lucas AM *et al.* applied the “CLeaning to Analysis: Reproducibility-based Interface for Traits and Exposures (CLARITE)” software to do quality control of the phenome by excluding low-quality phenotypes, thus preserving the power of PheWAS.[13] In particular, this software is user-customized and can be used for variable-specific QC. It has multiple functions such as separating qualitative and quantitative variables for QC, concurrently screening sample size minimums, identifying unique values, recoding missing values, and identifying documentation errors. Using CLARITE, Passero K *et al*. successfully selected phenotypes with a minimum sample size of 200, restricted QC to binary phenotypes, retained only samples which contained no missing covariate information and had at least a 99% sample call rate.[14] Detailed description and strengths and/or weaknesses of the software are summarized in **Supplementary Table 3.**

Step 2: Selecting predictors

Both genetic variants and non-genetic factors (e.g., serum biomarkers) can be used as predictors in PheWAS analysis. Utilizing SNPs derived from previous GWAS, PheWAS can detect novel associations or replicate known associations of a single or multiple variants with a variety of phenotypes. Denny JC *et al*. replicated 66% SNP-trait associations detected by GWAS and revealed 63 potentially pleiotropic associations with this strategy.[15] Functional genetic variants modifying the expression and/or activity of proteins are also used as genetic predictors to represent potential drug targets. For example, Jerome RN *et al*. conducted a PheWAS using genetic variants within *PCSK9* gene to explore any novel phenotype on which this protein and its inhibitors may have impact and thus to predict potential safety issues as well as side effects of drugs targeting on *PCSK9*.[16] Beyond using SNPs as genetic instruments individually, recent PheWASs have focused on using PRS as proxy of the exposure levels to explore their associations with a wide spectrum of phenotypes.[17-20] For instance, Li X *et al*. constructed a weighted PRS as a proxy of serum uric acid (SUA) levels in PheWAS analysis, and successfully identified significant associations with gout, hypertension and heart diseases.[19] Similarly, Meng X *et al*. created a PRS and implemented the PheWAS strategy to explore the role of vitamin D in health outcomes.[20] In addition, non-genetic predictors, such as autoantibodies status,[21] enzyme activity,[22] biomarker levels[23] and socio-economic factors[24] have also been explored in recent PheWASs to explore phenotypic associations.

Step 3: Phenotyping

Phenotyping refers to the process of defining an individual’s phenome. The phenome framework largely uses EMR-based binary disease phenotypes. For EMR-based PheWAS, the most straightforward way for phenotyping is to use the International Classification of Diseases (ICD) coding system. Two phenotyping methods are widely used to classify the ICD codes into appropriate case-control groups: the curated phenotyping and the holistic phenotyping. Curated phenotyping groups ICD codes that represent a common etiology into the same phenotype leading to a smaller number of phenotypes with more cases, thereby increasing statistical power and reducing the probability for false positive findings. Wu P *et al*. adopted the curated phenotyping strategy and aggregates ICD 10th Revision (ICD-10) codes into the PheCODE schema, which includes about 1,800 distinct phecodes.[25] By employing this schema, they successfully replicated several known genotype-phenotype associations with increased statistical power. In contrast, the holistic method tests all ICD codes and results in more phenotypes with relatively small number of cases. This method does not make any assumptions on the genetic or etiological commonalities of any disease, but the study power is largely reduced and results might be biased due to the correlations between phenotypes.[26] Cortes A *et al*. improved the holistic method by developing a tree-structured phenotypic model based on the hierarchical structure of ICD-10 codes and analyzed phenotypes by using a Bayesian analysis framework to increase the statistical power.[27] So far, the tree-structured phenotyping model has not been widely adopted. We only identified one study that used both the PheCODE schema and tree-structured phenotypic model, where the tree-structured phenotypic model showed advantages in detecting association with more sub-phenotypes.[19]

Phenomes can also be defined using data collected from epidemiological studies. With the increasing availability of large-scale-omics data, the possibility of exploring endo-phenotypes beyond binary clinical endpoints in PheWAS analyses has increased dramatically over the last few years. These could include laboratory biomarkers (proteomics, metabolomics or inflammatory biomarkers), anthropometric traits and many other phenotypes (e.g. socio-economic factors, imaging features). In addition, these intermediate phenotypes comprise largely quantitative traits, which could help detect earlier manifestations of diseases and enhance statistical power of PheWAS. Therefore, expanding the scope of the phenome offers an opportunity to detect associations with subclinical phenotypes. The first epidemiology-based PheWAS was conducted using the genomics and epidemiology (PAGE) network, in which related phenotypes were binned into the same class, and 111 significant phenotypic associations were finally identified.[28] Till present, there are a number of datasets with available clinical endpoints and sub-phenotypes that can be used to run a PheWAS analysis (**Table 2**). In epidemiology-based PheWAS, pooling epidemiological studies together is necessary in order to obtain a greater sample size and gain statistical power. However, this method may be hindered due to the difficulty to harmonize phenotypes between different studies, since the phenotype definition and the coverage of phenotypes may differ among studies.[29]

Step 4: Statistical analysis

For statistical analysis, linear regression for continuous variables and logistic regression for categorical variables are widely used to detect associations between predictors and disease outcomes with adjustment for a number of covariates. Generally, principal components (PCs) of ancestry are commonly included to reduce confounding by population structure. Demographic factors such as age and sex can influence the strength of genetic effects and therefore should also be adjusted for. But the selection of covariates should be cautious. The adjusted effect estimates may be less powerful with increasing stringency of Type 1 error control when the genetic variant is correlated with the covariates. Then, false discovery rate (FDR), permutation testing, and Bonferroni correction are appropriate ways to account for multiple testing. Till present, multiple strategies have been developed to perform PheWAS analysis. Traditional PheWAS using a single genetic variant lacks sufficient power in detecting phenotypic associations since a single genetic variant contributes small effects to several phenotypes or disorders. Thus, a novel PRS-PheWAS strategy has been proposed, in which a PRS is firstly calculated for each individual as the sum of risk increasing alleles weighted by the effect sizes taken from previous GWASs, then phenotypic associations between genetically determined exposure levels and health outcomes are examined by using the weighted PRS as a proxy. Leppert B *et al*. conducted a PheWAS analysis to examine associations of PRSs for five psychiatric disorders (major depression, bipolar disorder, schizophrenia, attention-deficit/hyperactivity disorder and autism spectrum disorder) with 23,004 outcomes in UK Biobank. The results showed 294 significant phenotypic associations, and most of them were related to mental health factors.[18] Another novel strategy, termed "MR-PheWAS", is to perform summary level data MR analyses across multiple phenotypes by using GWAS data to uncover the traits with potential causal associations. For example, Saunders *et al*. applied MR-PheWAS design to examine the causal associations between 316 intermediate phenotypes (which have GWAS summary data available from MR-Base platform) and glioma risk.[30] This approach opens up a new way to perform PheWAS when access to individual level data is not available, but it also has a number of methodological limitations. First, this approach is more like a candidate strategy rather than a phenome-wide test. Second, a distinct feature of PheWAS using individual level data is to examine the cross-phenotype associations in a single population, while in MR-PheWAS using GWAS data of multiple phenotypes, the ability to examine cross-phenotype associations will be dramatically diminished due to the substantial heterogeneity across different GWAS study populations. Additionally, an alternative Bayesian analysis has been developed for the tree-structured phenotypic model (referred as TreeWAS).[27, 31] In principle, it models the genetic coefficients across all phenotypes and a Markov process is applied to allow the genetic coefficients to evolve down the tree structure; Bayes factor statistic is calculated to evaluate the phenotypic associations. An example study using both PheWAS and TreeWAS analysis was performed by Li X *et al*. in which they identified several novel phenotypic associations in TreeWAS.[19]

Step 5: Interpretation of PheWAS results

Similar to GWAS, PheWAS could be regarded as a hypothesis generating analysis. When significant genotype-phenotype associations are identified, replication is needed to ensure that the positive results represent credible associations and are not chance findings or artifacts due to uncontrolled biases. Appropriate replication should be conducted in independent populations to validate the PheWAS findings. Furthermore, adjustment for specific covariates should be performed in sensitivity analyses to provide more accurate effect estimate. Finally, possible interpretations such as causality, pleiotropy, true comorbidity or confounded phenotype relationships should be taken into account to understand the PheWAS associations. Currently, a novel strategy that firstly performs a PheWAS analysis and then applies MR to validate significant PheWAS findings has become an efficient way to explore the causal effects of an exposure (e.g., specific biomarker) on a large number of phenotypes. For example, Li X *et al*. performed a PheWAS to identify disease outcomes associated with genetic risk loci of SUA level and then implemented MR analysis to investigate the causal relevance between SUA level and identified disease outcomes.[32] MendelianRandomization[33] and TwoSampleMR[34] are commonly used packages. Additionally, generalized summary statistic based Mendelian randomization (GSMR) was developed to perform bi-directional MR analysis and distinguish genetic pleiotropy from genetic linkage using the heterogeneity in dependent instrument (HEIDI) test.[35] For example, Li X *et al*. investigated the phenotypic associations of multiple SUA genetic loci and applied the HEIDI approach to identify independent causal variants that affect multiple health conditions.[32]

Step 6: Visualization of PheWAS results

We identified a number of useful tools that could be used to graphically represent PheWAS results, such as PheWAS-View[36] and PhenoGram.[37] Manhattan plot and heatmap are commonly used to visualize PheWAS results, and they could be generated by the PheWAS-View tool. Manhattan plot can be used to visualize all association results across phenotypes and heatmap plot can distinguish between correlated phenotypes and possible pleiotropy by taking all pairwise correlation between coefficients into account (**Supplementary Figure 1A-1B**). PhenoGram is another web-based tool that can be used to visualize the genomic or single chromosomal coverage of SNPs, and the plot can display the shared genomic information among different phenotypes (**Supplementary Figure 1C**). Both PheWAS-View and PhenoGram have been widely employed to visually integrate PheWAS results in the published studies.[28, 38-45]

Online resources and tools for PheWAS analysis

Multiple alternative approaches used for PheWAS analysis have been developed. PheWAS package is a commonly used software to perform PheWAS analysis by integrating ICD-9 and ICD-10 codes. Additionally, SPAtest[46] and SAIGE[47] are options which are more computationally efficient and scalable to PheWAS analysis in large cohorts and biobanks by using the saddlepoint approximation (SPA) method, a powerful tool for obtaining accurate expressions for densities and distribution functions.[48, 49] In term of its application to PheWAS, SPA has the ability to correct the inflated type I error caused by the unbalanced case-control ratio through adjusting single-variant score statistics and is faster than other existing rare-variant tests. Based on the availability of multi-omic data such as genotype markers (genome), gene expression measurements (transcriptome) and clinical traits measurements (phenome), BioBin[50] and GenAMap[51] were developed to enable the identification of the biological mechanisms underlying significant PheWAS associations. In particular, BioBin can be applied for PheWAS analysis of rare genetic variant to enhance study power. Similarly, PLATO[52] is suitable for analyzing rare variant PheWAS and also applicable to analyze non-genetic data such as environmental data. Moreover, PHESANT is another tool made available to test the association of a specified trait with all variables in UK Biobank.[53] All these afore mentioned software have their own strengths and weaknesses. The most appropriate tool could be applied based on the motivation for performing PheWAS as well as the data structure. For example, researchers can apply BioBin, SPAtest and SAIGE to increase the power of PheWAS for binary phenotypes by dealing with extremely unbalanced case-control ratios or rare outcomes in large biobank data.

By taking advantage of the wide range of publicly available datasets, a number of large integrated databases and resources are made available to implement PheWAS analysis (**Table 3**). PheWAS catalog is a publicly available resource of a large-scale PheWAS testing for the associations of SNPs derived from the NHGRI-EBI GWAS catalog with a spectrum of phenotypes.[15] Similarly, PhenoScanner is a curated database of publicly available results from genetic association studies, aiming to facilitate phenome scans and provide insights for the understanding of disease pathways and biology.[54] Other platforms such as MR-Base[34] and GeneATLAS[55] contain millions of associations between genetic variants and traits identified from previous GWASs and also allow the potential phenotypic relationships to be efficiently evaluated in PheWAS analysis.

Discussion

We conducted a systematic review of all published PheWAS studies and summarized the main steps for performing a PheWAS as well as the characteristics of commonly used software for each step. The strengths, challenges, and potential implications for future PheWAS are further discussed here.

Advantages of PheWAS

PheWAS is suggested to be an efficient way to test for novel genotype-phenotype associations. A unique advantage of PheWAS is that it has the ability to explore associations of genetic predictors with a wide spectrum of phenotypes simultaneously. Besides, among the identified phenotypic associations, multiple genetic variants are found to be significantly associated with two or more phenotypes, indicating the capacity of PheWAS in detecting pleiotropic variants.[15] The identification of genetic pleiotropy provides significant insight into the shared genetic etiologies of multiple diseases. For example, Zheutlin AB *et al*. identified that the PRS of schizophrenia was also associated with other psychiatric disorders such as anxiety, neurological and personality disorders, suicidal behavior and memory loss, indicating shared genetic risk among these phenotypes.[56] PheWAS also serves as a useful tool to identify potential causal relationships across phenotypes. Focusing on variants that have known biological function and/or clinical significance, PheWAS has the capacity to simplify the interpretation of novel PheWAS results by directly applying the background information of genetic variants and thus uncover potential causality.[26] For example, Salem JE *et al*. created a PRS using genetic variants related to thyrotropin levels and found a causal relationship between thyrotropin levels and atrial fibrillation.[57]

Challenges and future directions for PheWAS

There are several challenges in the current scope of PheWAS. Most PheWASs focus on SNPs identified by GWAS. However, GWAS findings only account for a small proportion of the genetic variance of a specific biomarker and the majority of the GWAS findings do not represent functional variants, resulting in a big amount of information that could be potentially informative underutilized. An alternate PheWAS approach could be to focus on candidate SNPs or variants derived from next generation sequencing (NGS), which captures a great deal of genetic data across human genome including indels, complex rearrangements, copy number variants and many rare variants.[58] This strategy allows a PheWAS for every SNP or variant across the genome, which can be regarded as a complement to the current PheWAS-by-GWAS approach. Furthermore, non-genetic variables can also serve as the focus of PheWAS. Liao KP *et al*. studied the association between rheumatoid arthritis (RA) related autoantibodies levels and multiple clinical phenotypes.[59] Cai W *et al*. applied PheWAS to explore the association of health care costs with inflammatory bowel diseases.[24]

Selection of appropriate covariates for adjustment is an important issue in PheWAS in order to derive unbiased estimates. The PheWAS design uses genetic variants as instruments (IVs) to assess the influence of modifiable exposures on a wide range of health outcomes. As germline genetic variants are generally independent of confounding factors and are determined at conception, adjustment for a wide range of traditional confounders is usually not suggested. However, in some cases, adjustment for covariates is necessary to ensure validity of the IVs, as the IV assumptions affect only a subgroup of interest. An example is the case of population stratification, in which the sample population consists of subpopulations (e.g., ethnic subgroups) that have different distributions of the IVs and outcome. Association between the IVs and outcome may solely correspond to differences in ethnicity and not to any biological effects of the exposure. This can be addressed at least partially by adjusting for genetic PCs. Additionally, if there are measured covariates, which explain variation in the exposure or a continuous outcome, and which are not on the causal pathway between exposure and outcome, then they could be adjusted as covariates to increase the statistical power. However, it should be done with cautions as this may lead to collider bias when a covariate is on the causal pathway between exposure and outcome or causally downstream of a collider.[60] As PheWAS examines the associations with a wide range of phenotypes, it is impractical to assess the risk of collider bias for all disease outcomes. Therefore, adjustment for additional covariates is generally discouraged in the primary analysis. Instead, a sensitivity analysis after PheWAS to adjust for covariates specific to the identified genotype-phenotype association of interest should be performed to complement the interpretation of the PheWAS results.

As we are moving towards genome-wide PheWAS, the number of genetic variants involved in the PheWAS analysis will undoubtedly grow. In addition, as the collection of patient health data increases, more phenotypes will become available. Therefore, PheWAS could be challenged by a growing multiple comparison burden. Common methods such as Bonferroni correction can be used for correcting for multiple testing. However, due to the large number of phenotypes involved in the PheWAS analysis as well as the inter-correlations between phenotypes, Bonferroni method is overly strict. Although FDR, inter-data replication and permutation perform better in calculating the pairwise correlation between the phenotypes, still more advanced methods need to be developed in the future to correct more efficiently for multiple tests. Notably, both type 1 error rate (false positives) and type 2 error rate (false negatives) should be controlled in the future development of robust strategies for dealing with multiple testing.[61]

Current EMR-based PheWAS applies an automatic phenotyping algorithm based on ICD codes. Although using ICD codes to define the phenome is efficient and cost-effective, it also has some disadvantages. It has already been noted that phenotyping based on ICD coding can lead to an increased number of false positive in which billing codes do not represent medical conditions.[4] To address this, specificity can be increased by adopting rigid threshold approaches such as the ‘rule-of-two’, which means that cases are defined when there are at least two instances of that code in their records.[62] However, the sensitivity (detection rate) may also fall thus reducing the number of cases, which could affect the power of the study. Besides, phenotyping using thresholding methods does not take total health care utilization (i.e. total number of the billing codes for a specific phenotype) into consideration, resulting in different probabilities to be a case despite the numbers of diagnosis codes are the same. Thus, more advanced phenotyping methods are warranted to be established for future PheWAS studies. An improvement has been made by PheProb approach which clusters patients into two groups (likely cases and likely non-cases) based on the probability of being a case calculated using the number of billing codes and total health care utilization.[63] EMR-based PheWAS is limited to binary disease phenotypes, future efforts should be made to expand the phenotypic domains by including other quantitative traits such as laboratory values, body measurements or imaging data. For example, Córdova-Palomera A *et al*. created a PRS for aortic valve area using genotyping data and aortic valve area measurements from magnetic resonance imaging (MRI) sequence data in UK Biobank and then performed a PheWAS analysis to identify genetic comorbidities.[64] This study illustrated the use of automated phenotyping of cardiac imaging data to investigate the genetic etiology of aortic valve diseases, uncover the correlations between genetic factors and cardiac anatomy, and guide clinical diagnosis and prediction. Thus, expanding the scope of PheWAS analysis to include quantitative traits offers an opportunity to detect associations with subclinical phenotypes.

Although identification of cross-phenotype effects is a strength of PheWAS, the biggest challenge is the interpretation of the observed associations. It is challenging to identify the biological mechanisms and clinical significance behind PheWAS associations when utilizing GWAS SNPs, which are predominantly tag SNPs and reside primarily in intergenic regions with unknown function. In addition, novel bioinformatics methods are warranted to be developed to conduct a series of post-PheWAS analyses to functionally characterize the variants. Firstly, fine mapping is helpful to reveal the causal variants that are in linkage disequilibrium (LD) with the markers.[65] Then, functional annotation of the causal variants (gene location, missense or nonsense) based on public available databases such as PolyPhen[66] and SIFT[67] can provide important insights for biological function. Furthermore, quantitative trait loci (QTL) analyses on multi-omic data including gene expression, protein activity and metabolite level can help identify the regulatory changes caused by mutations.[68, 69]. The identified cross-phenotype effects can also occur when the associated gene is involved in different biological processes or only one pathway that has diverse effects on multiple phenotypes. In this case, pathway enrichment analysis can be carried out using public resources of pathways such as Gene Oncology (GO)[70] and Kyoto Encyclopedia of Genes and Genomes (KEGG)[71] to search for the biological connections among phenotypes.

Clinical applications of PheWAS

PheWAS has been recognized as a promising approach to establish novel treatment strategies through drug repositioning, which refers to the application of an existing therapeutic drug for new indications that share common pathophysiology.[72, 73] Millwood IY *et al*. performed a PheWAS to identify genetic associations of a functional variant that inactivates lipoprotein-associated phospholipase A2 (Lp-PLA2) activity with a wide range of disease outcomes. They found that lifelong lower Lp-PLA2 activity was not associated with major risks of vascular or non-vascular diseases.[74] This finding challenges the protective role of Lp-PLA2 inhibitor like darapladib in preventing major vascular diseases and provides important insights for further drug development. Moreover, Jerome RN *et al.* examined PheWAS data for 16 genes and detected therapeutic indications for 13 of 16 gene-targeted drugs.[75] Beyond repositioning, PheWAS approach could be used to predict potential side effects associated with drug use, which prompts drug development in the early stages of clinical trials.[76] So far, the PheWAS design has successfully elucidate the possible efficacy and adverse effects of antihypertensive drug,[77] lipid-lowering drug [78] and antidepressant drug.[38]

Conclusions

In summary, PheWAS provides an efficient way to identify phenotypic associations in a high-throughput manner. PheWAS methods and associated software have evolved rapidly, although several challenges still remain to be overcome. In the future, even larger populations and more diverse data types will be involved in PheWAS analysis, which leads to a need for more optimized methods to be created. Applying biological knowledge in a standardized framework to aid the interpretation of PheWAS results is an essential part of future studies. Application of PheWAS approach in drug repurposing remains a research focus to prompt drug development.

Abbreviations

eMERGE Network: Electronic medical records and genomics network
EMR: Electronic medical records
FDR: False discovery rate
GO: Gene oncology
PRS: Polygenic risk scores
GSMR: Generalized summary statistic based mendelian randomization
GWAS: Genome wide association study
HEIDI: Heterogeneity in dependent instrument
ICD: International classification of diseases
KEGG: Kyoto encyclopedia of genes and genomes

LD: Linkage disequilibrium
MR: Mendelian randomization
MVP: Million veteran program
NGS: Next generation sequencing
PheWAS: Phenome-wide association study
QC: Quality control
QTL: Quantitative trait loci
SNP: Single nucleotide polymorphism
TreeWAS: Tree-structured phenotypic model

Acknowledgements

We thank all study participants for their contributions to this study.

Competing interests

The authors declare no competing interests.

Funding

This work was supported by funding for the infrastructure and staffing of the Edinburgh CRUK Cancer Research Centre. E.T. is supported by a CRUK Career Development Fellowship (C31250/A22804).

Authors' contributions

Lijuan Wang: Conceptualization, Literature review, Data extraction, Writing-original draft.
Xiaomeng Zhang: Conceptualization, Writing-review & editing.
Xiangrui Meng: Conceptualization, Writing-review & editing.
Fotios Koskeridis: Literature review, Data extraction, Writing-review & editing.
Andrea Georgiou: Literature review, Data extraction, Writing-review & editing.
Lili Yu: Data extraction.
Harry Campbell: Conceptualization, Writing-review & editing.

Evropi Theodoratou: Conceptualization, Supervision, Writing-review & editing.

Xue Li: Conceptualization, Supervision, Writing-review & editing.

All authors read and approved the final manuscript.

Table 1. Characteristics of eligible studies.

Characteristics	Number of studies (%)
Sample size (n = 178)	
Small (< 1,000 subjects)	5 (2.8)
Large (1,000-9,999 subjects)	36 (20.2)
Very Large (⩾ 10,000 subjects)	137 (77.0)
Definition of phenome (n = 178)	
EMR-based phenome	164 (92.1)
Epidemiology-based phenome	14 (7.9)
Predictor (n = 178)	
Single SNP	21 (11.8)
Multiple SNPs were used to construct Polygenic risk score (PRS)	31 (17.4)
Multiple SNPs	88 (49.4)
Biomarker	10 (5.6)
Other	28 (15.7)
Conducted replication analysis (n = 178)	
Yes	73 (41.0)
Validated PheWAS conclusion	69 (94.5)
Did not validate PheWAS conclusion	4 (5.5)
No	105 (59.0)

Note: 17 papers introducing novel software were not included in this summary.

Table 2. Datasets used for PheWAS analysis.

Datasets	Genotype data	Clinical endpoints	Laboratory biomarkers	Physical measurements	Socio-economic factors
AIDS Clinical Trials Group (ACTG) Network	√	√	√		
Atherosclerosis Risk in Communities (ARIC)	√	√	√	√	√
Avon Longitudinal Study of Parents and Children (ALSPAC)	√	√	√	√	√
BioBank Japan Project (BBJ)	√	√	√	√	√
BioMe Biobank	√	√	√	√	√
China Kadoorie Biobank (CKB)	√	√	√	√	√
Electronic Medical Records and Genomics (eMERGE) Network	√	√	√	√	√
Estonian Biobank	√	√	√	√	√
Genomics Evidence Neoplasia Information Exchange (GENIE)	√	√			
Global Lipids Genetics Consortium (GLGC)	√	√	√		
Integrative Psychiatric Research (iPSYCH)	√	√	√	√	√
Lifelines cohort	√	√	√	√	√
Long Life Family Study (LLFS)	√	√	√	√	√
Mass General Brigham Biobank (MGBB)	√	√	√	√	√
Michigan Genomics Initiative (MGI)	√	√	√	√	
Million Veteran Program (MVP)	√	√	√	√	√
Northern Nevada Cohort	√	√			
Population Architecture using Genomics and Epidemiology (PAGE) Network	√	√	√	√	√
Penn Medicine Biobank (PMBB)	√	√	√		

Twins Early Development Study (TEDS)	√	√	√	√	√
UK Biobank (UKB)	√	√	√	√	√
Women’s Health Initiative (WHI)	√	√	√	√	√

Table 3. Software integrating databases and resources for PheWAS analysis.

Software	Description	Reference
PheWAS catalogue	A publicly available resource of PheWASs and their results.	Denny JC et al., 2013 [15]
PheKB	An online collaborative environment supporting the workflow of building, sharing, and validating electronic phenotype algorithms.	Kirby JC et al., 2016 [79]
PhenoScanner	A curated database of publicly available results from human genotype-phenotype association studies.	Staley JR et al., 2016 [54]
MR-Base	A platform for Mendelian randomization and PheWAS by integrates a curated database of GWAS summary data, enabling millions of potential causal relationships to be systematically and efficiently evaluated in PheWAS.	Hemani G et al., 2018 [34]
GeneATLAS	A large database of associations between hundreds of traits and millions of variants using the UK Biobank cohort.	Canela-Xandri O et al., 2018 [55]

Figure legends

Figure 1. Flow chart of the study selection process of the systematic literature review of PheWAS studies.

Figure 2. Flow chart of main steps to conduct PheWAS analysis. SNP: single nucleotide polymorphisms; PRS: polygenic risk score; EMR: electronic medical record; ICD: international classification of diseases; MR: mendelian randomization; IVW: inverse-variance weighted; FDR: false discovery rate.

References:

1. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet 2017;**101**(1):5-22 doi: 10.1016/j.ajhg.2017.06.005[published Online First: Epub Date]|.
2. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, Pendlington ZM, Welter D, Burdett T, Hindorff L, Flicek P, Cunningham F, Parkinson H. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res 2017;**45**(D1):D896-D901 doi: 10.1093/nar/gkw1133[published Online First: Epub Date]|.
3. Tam V, Patel N, Turcotte M, Bosse Y, Pare G, Meyre D. Benefits and limitations of genome-wide association studies. Nat Rev Genet 2019;**20**(8):467-84 doi: 10.1038/s41576-019-0127-1[published Online First: Epub Date]|.
4. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, Wang D, Masys DR, Roden DM, Crawford DC. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics 2010;**26**(9):1205-10 doi: 10.1093/bioinformatics/btq126[published Online First: Epub Date]|.
5. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012;**13**(6):395-405 doi: 10.1038/nrg3208[published Online First: Epub Date]|.
6. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, Cortes A, Welsh S, Young A, Effingham M, McVean G, Leslie S, Allen N, Donnelly P, Marchini J. The UK Biobank resource with deep phenotyping and genomic data. Nature 2018;**562**(7726):203-09 doi: 10.1038/s41586-018-0579-z[published Online First: Epub Date]|.
7. Chen Z, Chen J, Collins R, Guo Y, Peto R, Wu F, Li L, China Kadoorie Biobank collaborative g. China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. Int J Epidemiol 2011;**40**(6):1652-66 doi: 10.1093/ije/dyr120[published Online First: Epub Date]|.
8. 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, Bottinger EP, Williams MS, e MN. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med 2013;**15**(10):761-71 doi: 10.1038/gim.2013.72[published Online First: Epub Date]|.
9. Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, Breeling J, Whitbourne S, Deen J, Shannon C, Humphries D, Guarino P, Aslan M, Anderson D, LaFleur R, Hammond T, Schaa K, Moser J, Huang G, Muralidhar S, Przygodzki R, O'Leary TJ. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. J Clin Epidemiol 2016;**70**:214-23 doi: 10.1016/j.jclinepi.2015.09.016[published Online First: Epub Date]|.
10. Bush WS, Oetjens MT, Crawford DC. Unravelling the human genome-phenome relationship using phenome-wide association studies. Nat Rev Genet 2016;**17**(3):129-45 doi: 10.1038/nrg.2015.36[published Online First: Epub Date]|.
11. Verma A, Bradford Y, Dudek S, Lucas AM, Verma SS, Pendergrass SA, Ritchie MD. A simulation study investigating power estimates in phenome-wide association studies. BMC Bioinformatics 2018;**19**(1):120 doi: 10.1186/s12859-018-2135-0[published Online First: Epub Date]|.

12. Namjou B, Marsolo K, Caroll RJ, Denny JC, Ritchie MD, Verma SS, Lingren T, Porollo A, Cobb BL, Perry C, Kottyan LC, Rothenberg ME, Thompson SD, Holm IA, Kohane IS, Harley JB. Phenome-wide association study (PheWAS) in EMR-linked pediatric cohorts, genetically links PLCL1 to speech language development and IL5-IL13 to Eosinophilic Esophagitis. Front Genet 2014;**5**:401 doi: 10.3389/fgene.2014.00401[published Online First: Epub Date]|.

13. Lucas AM, Palmiero NE, McGuigan J, Passero K, Zhou J, Orie D, Ritchie MD, Hall MA. CLARITE Facilitates the Quality Control and Analysis Process for EWAS of Metabolic-Related Traits. Front Genet 2019;**10**:1240 doi: 10.3389/fgene.2019.01240[published Online First: Epub Date]|.

14. Passero K, He X, Zhou J, Mueller-Myhsok B, Kleber ME, Maerz W, Hall MA. Phenome-wide association studies on cardiovascular health and fatty acids considering phenotype quality control practices for epidemiological data. Pac Symp Biocomput 2020;**25**:659-70

15. 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**(12):1102-10 doi: 10.1038/nbt.2749[published Online First: Epub Date]|.

16. Jerome RN, Pulley JM, Roden DM, Shirey-Rice JK, Bastarache LA, G RB, L BE, Lancaster WJ, Denny JC. Using Human 'Experiments of Nature' to Predict Drug Safety Issues: An Example with PCSK9 Inhibitors. Drug Saf 2018;**41**(3):303-11 doi: 10.1007/s40264-017-0616-0[published Online First: Epub Date]|.

17. Fritsche LG, Gruber SB, Wu Z, Schmidt EM, Zawistowski M, Moser SE, Blanc VM, Brummett CM, Kheterpal S, Abecasis GR, Mukherjee B. Association of Polygenic Risk Scores for Multiple Cancers in a Phenome-wide Study: Results from The Michigan Genomics Initiative. Am J Hum Genet 2018;**102**(6):1048-61 doi: 10.1016/j.ajhg.2018.04.001[published Online First: Epub Date]|.

18. Leppert B, Millard LAC, Riglin L, Davey Smith G, Thapar A, Tilling K, Walton E, Stergiakouli E. A cross-disorder PRS-pheWAS of 5 major psychiatric disorders in UK Biobank. PLoS Genet 2020;**16**(5):e1008185 doi: 10.1371/journal.pgen.1008185[published Online First: Epub Date]|.

19. Li X, Meng X, He Y, Spiliopoulou A, Timofeeva M, Wei WQ, Gifford A, Yang T, Varley T, Tzoulaki I, Joshi P, Denny JC, McKeigue P, Campbell H, Theodoratou E. Genetically determined serum urate levels and cardiovascular and other diseases in UK Biobank cohort: A phenome-wide mendelian randomization study. PLoS Med 2019;**16**(10):e1002937 doi: 10.1371/journal.pmed.1002937[published Online First: Epub Date]|.

20. Meng X, Li X, Timofeeva MN, He Y, Spiliopoulou A, Wei WQ, Gifford A, Wu H, Varley T, Joshi P, Denny JC, Farrington SM, Zgaga L, Dunlop MG, McKeigue P, Campbell H, Theodoratou E. Phenome-wide Mendelian-randomization study of genetically determined vitamin D on multiple health outcomes using the UK Biobank study. Int J Epidemiol 2019;**48**(5):1425-34 doi: 10.1093/ije/dyz182[published Online First: Epub Date]|.

21. Liao KP, Kurreeman F, Li G, Duclos G, Murphy S, Guzman R, Cai T, Gupta N, Gainer V, Schur P, Cui J, Denny JC, Szolovits P, Churchill S, Kohane I, Karlson EW, Plenge RM. Associations of autoantibodies, autoimmune risk alleles, and clinical diagnoses from the electronic medical records in rheumatoid

arthritis cases and non-rheumatoid arthritis controls. Arthritis Rheum 2013;**65**(3):571-81 doi: 10.1002/art.37801[published Online First: Epub Date]|.

22. Neuraz A, Chouchana L, Malamut G, Le Beller C, Roche D, Beaune P, Degoulet P, Burgun A, Loriot MA, Avillach P. Phenome-wide association studies on a quantitative trait: application to TPMT enzyme activity and thiopurine therapy in pharmacogenomics. PLoS Comput Biol 2013;**9**(12):e1003405 doi: 10.1371/journal.pcbi.1003405[published Online First: Epub Date]|.
23. Feng Q, Wei WQ, Chung CP, Levinson RT, Sundermann AC, Mosley JD, Bastarache L, Ferguson JF, Cox NJ, Roden DM, Denny JC, Linton MF, Edwards DRV, Stein CM. Relationship between very low low-density lipoprotein cholesterol concentrations not due to statin therapy and risk of type 2 diabetes: A US-based cross-sectional observational study using electronic health records. PLoS Med 2018;**15**(8):e1002642 doi: 10.1371/journal.pmed.1002642[published Online First: Epub Date]|.
24. Cai W, Cagan A, He Z, Ananthakrishnan AN. A Phenome-Wide Analysis of Healthcare Costs Associated with Inflammatory Bowel Diseases. Dig Dis Sci 2020 doi: 10.1007/s10620-020-06329-9[published Online First: Epub Date]|.
25. Wu P, Gifford A, Meng X, Li X, Campbell H, Varley T, Zhao J, Carroll R, Bastarache L, Denny JC, Theodoratou E, Wei WQ. Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. JMIR Med Inform 2019;**7**(4):e14325 doi: 10.2196/14325[published Online First: Epub Date]|.
26. Hebbring SJ. The challenges, advantages and future of phenome-wide association studies. Immunology 2014;**141**(2):157-65 doi: 10.1111/imm.12195[published Online First: Epub Date]|.
27. Cortes A, Dendrou CA, Motyer A, Jostins L, Vukcevic D, Dilthey A, Donnelly P, Leslie S, Fugger L, McVean G. Bayesian analysis of genetic association across tree-structured routine healthcare data in the UK Biobank. Nat Genet 2017;**49**(9):1311-18 doi: 10.1038/ng.3926[published Online First: Epub Date]|.
28. Pendergrass SA, Brown-Gentry K, Dudek S, Frase A, Torstenson ES, Goodloe R, Ambite JL, Avery CL, Buyske S, Buzkova P, Deelman E, Fesinmeyer MD, Haiman CA, Heiss G, Hindorff LA, Hsu CN, Jackson RD, Kooperberg C, Le Marchand L, Lin Y, Matise TC, Monroe KR, Moreland L, Park SL, Reiner A, Wallace R, Wilkens LR, Crawford DC, Ritchie MD. 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 doi: 10.1371/journal.pgen.1003087[published Online First: Epub Date]|.
29. Evangelou E, Ioannidis JP. Meta-analysis methods for genome-wide association studies and beyond. Nat Rev Genet 2013;**14**(6):379-89 doi: 10.1038/nrg3472[published Online First: Epub Date]|.
30. Saunders CN, Cornish AJ, Kinnersley B, Law PJ, Houlston RS, Collaborators. Searching for causal relationships of glioma: a phenome-wide Mendelian randomisation study. Br J Cancer 2021;**124**(2):447-54 doi: 10.1038/s41416-020-01083-1[published Online First: Epub Date]|.
31. Cox NJ. Reaching for the next branch on the biobank tree of knowledge. Nat Genet 2017;**49**(9):1295-96 doi: 10.1038/ng.3946[published Online First: Epub Date]|.
32. Li X, Meng X, Spiliopoulou A, Timofeeva M, Wei WQ, Gifford A, Shen X, He Y, Varley T, McKeigue P, Tzoulaki I, Wright AF, Joshi P, Denny JC, Campbell H, Theodoratou E. MR-PheWAS: exploring the causal effect of SUA level on multiple disease outcomes by using genetic instruments in UK Biobank. Ann Rheum Dis 2018;**77**(7):1039-47 doi: 10.1136/annrheumdis-2017-212534[published Online First: Epub Date]|.

33. Yavorska OO, Burgess S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. Int J Epidemiol 2017;**46**(6):1734-39 doi: 10.1093/ije/dyx034[published Online First: Epub Date]|.

34. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, Laurin C, Burgess S, Bowden J, Langdon R, Tan VY, Yarmolinsky J, Shihab HA, Timpson NJ, Evans DM, Relton C, Martin RM, Davey Smith G, Gaunt TR, Haycock PC. The MR-Base platform supports systematic causal inference across the human phenome. Elife 2018;**7** doi: 10.7554/eLife.34408[published Online First: Epub Date]|.

35. Zhu Z, Zheng Z, Zhang F, Wu Y, Trzaskowski M, Maier R, Robinson MR, McGrath JJ, Visscher PM, Wray NR, Yang J. Causal associations between risk factors and common diseases inferred from GWAS summary data. Nat Commun 2018;**9**(1):224 doi: 10.1038/s41467-017-02317-2[published Online First: Epub Date]|.

36. 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 doi: 10.1186/1756-0381-5-5[published Online First: Epub Date]|.

37. Wolfe D, Dudek S, Ritchie MD, Pendergrass SA. Visualizing genomic information across chromosomes with PhenoGram. BioData Min 2013;**6**(1):18 doi: 10.1186/1756-0381-6-18[published Online First: Epub Date]|.

38. Verma SS, Josyula N, Verma A, Zhang X, Veturi Y, Dewey FE, Hartzel DN, Lavage DR, Leader J, Ritchie MD, Pendergrass SA. Rare variants in drug target genes contributing to complex diseases, phenome-wide. Sci Rep 2018;**8**(1):4624 doi: 10.1038/s41598-018-22834-4[published Online First: Epub Date]|.

39. Carroll RJ, Bastarache L, Denny JC. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. Bioinformatics 2014;**30**(16):2375-6 doi: 10.1093/bioinformatics/btu197[published Online First: Epub Date]|.

40. Oetjens MT, Bush WS, Denny JC, Birdwell K, Kodaman N, Verma A, Dilks HH, Pendergrass SA, Ritchie MD, Crawford DC. Evidence for extensive pleiotropy among pharmacogenes. Pharmacogenomics 2016;**17**(8):853-66 doi: 10.2217/pgs-2015-0007[published Online First: Epub Date]|.

41. Hall MA, Verma A, Brown-Gentry KD, Goodloe R, Boston J, Wilson S, McClellan B, Sutcliffe C, Dilks HH, Gillani NB, Jin H, Mayo P, Allen M, Schnetz-Boutaud N, Crawford DC, Ritchie MD, Pendergrass SA. 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. PLoS Genet 2014;**10**(12):e1004678 doi: 10.1371/journal.pgen.1004678[published Online First: Epub Date]|.

42. Verma A, Basile AO, Bradford Y, Kuivaniemi H, Tromp G, Carey D, Gerhard GS, Crowe JE, Jr., Ritchie MD, Pendergrass SA. Phenome-Wide Association Study to Explore Relationships between Immune System Related Genetic Loci and Complex Traits and Diseases. PLoS One 2016;**11**(8):e0160573 doi: 10.1371/journal.pone.0160573[published Online First: Epub Date]|.

43. Verma SS, Lucas AM, Lavage DR, Leader JB, Metpally R, Krishnamurthy S, Dewey F, Borecki I, Lopez A, Overton J, Penn J, Reid J, Pendergrass SA, Breitwieser G, Ritchie MD. Identifying Genetic Associations with Variability in Metabolic Health and Blood Count Laboratory Values: Diving into the Quantitative Traits by Leveraging Longitudinal Data from an Ehr. Pac Symp Biocomput 2017;**22**:533-44 doi: 10.1142/9789813207813_0049[published Online First: Epub Date]|.

44. Moore CB, Verma A, Pendergrass S, Verma SS, Johnson DH, Daar ES, Gulick RM, Haubrich R, Robbins GK, Ritchie MD, Haas DW. Phenome-wide Association Study Relating Pretreatment Laboratory

Parameters With Human Genetic Variants in AIDS Clinical Trials Group Protocols. Open Forum Infect Dis 2015;**2**(1):ofu113 doi: 10.1093/ofid/ofu113[published Online First: Epub Date]|.

45. Pendergrass SA, Buyske S, Jeff JM, Frase A, Dudek S, Bradford Y, Ambite JL, Avery CL, Buzkova P, Deelman E, Fesinmeyer MD, Haiman C, Heiss G, Hindorff LA, Hsu CN, Jackson RD, Lin Y, Le Marchand L, Matise TC, Monroe KR, Moreland L, North KE, Park SL, Reiner A, Wallace R, Wilkens LR, Kooperberg C, Ritchie MD, Crawford DC. A phenome-wide association study (PheWAS) in the Population Architecture using Genomics and Epidemiology (PAGE) study reveals potential pleiotropy in African Americans. PLoS One 2019;**14**(12):e0226771 doi: 10.1371/journal.pone.0226771[published Online First: Epub Date]|.
46. Dey R, Schmidt EM, Abecasis GR, Lee S. A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. Am J Hum Genet 2017;**101**(1):37-49 doi: 10.1016/j.ajhg.2017.05.014[published Online First: Epub Date]|.
47. Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, LeFaive J, VandeHaar P, Gagliano SA, Gifford A, Bastarache LA, Wei WQ, Denny JC, Lin M, Hveem K, Kang HM, Abecasis GR, Willer CJ, Lee S. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet 2018;**50**(9):1335-41 doi: 10.1038/s41588-018-0184-y[published Online First: Epub Date]|.
48. Kuonen D. Saddlepoint approximations for distributions of quadratic forms in normal variables. Biometrika 1999;**86**(4):929-35 doi: 10.1093/biomet/86.4.929[published Online First: Epub Date]|.
49. Daniels HE. Saddlepoint Approximations in Statistics. Ann Math Stat 1954;**25**(4):631-50
50. Moore CB, Wallace JR, Frase AT, Pendergrass SA, Ritchie MD. BioBin: a bioinformatics tool for automating the binning of rare variants using publicly available biological knowledge. BMC Med Genomics 2013;**6 Suppl 2**:S6 doi: 10.1186/1755-8794-6-S2-S6[published Online First: Epub Date]|.
51. Xing EP, Curtis RE, Schoenherr G, Lee S, Yin J, Puniyani K, Wu W, Kinnaird P. GWAS in a box: statistical and visual analytics of structured associations via GenAMap. PLoS One 2014;**9**(6):e97524 doi: 10.1371/journal.pone.0097524[published Online First: Epub Date]|.
52. Hall MA, Wallace J, Lucas A, Kim D, Basile AO, Verma SS, McCarty CA, Brilliant MH, Peissig PL, Kitchner TE, Verma A, Pendergrass SA, Dudek SM, Moore JH, Ritchie MD. PLATO software provides analytic framework for investigating complexity beyond genome-wide association studies. Nat Commun 2017;**8**(1):1167 doi: 10.1038/s41467-017-00802-2[published Online First: Epub Date]|.
53. Millard LAC, Davies NM, Gaunt TR, Davey Smith G, Tilling K. Software Application Profile: PHESANT: a tool for performing automated phenome scans in UK Biobank. Int J Epidemiol 2018;**47**(1):29-35 doi: 10.1093/ije/dyx204[published Online First: Epub Date]|.
54. Staley JR, Blackshaw J, Kamat MA, Ellis S, Surendran P, Sun BB, Paul DS, Freitag D, Burgess S, Danesh J, Young R, Butterworth AS. PhenoScanner: a database of human genotype-phenotype associations. Bioinformatics 2016;**32**(20):3207-09 doi: 10.1093/bioinformatics/btw373[published Online First: Epub Date]|.
55. Canela-Xandri O, Rawlik K, Tenesa A. An atlas of genetic associations in UK Biobank. Nat Genet 2018;**50**(11):1593-99 doi: 10.1038/s41588-018-0248-z[published Online First: Epub Date]|.
56. Zheutlin AB, Dennis J, Karlsson Linner R, Moscati A, Restrepo N, Straub P, Ruderfer D, Castro VM, Chen CY, Ge T, Huckins LM, Charney A, Kirchner HL, Stahl EA, Chabris CF, Davis LK, Smoller JW. Penetrance and Pleiotropy of Polygenic Risk Scores for Schizophrenia in 106,160 Patients Across Four

Health Care Systems. Am J Psychiatry 2019;**176**(10):846-55 doi: 10.1176/appi.ajp.2019.18091085[published Online First: Epub Date]|.

57. Salem JE, Shoemaker MB, Bastarache L, Shaffer CM, Glazer AM, Kroncke B, Wells QS, Shi M, Straub P, Jarvik GP, Larson EB, Velez Edwards DR, Edwards TL, Davis LK, Hakonarson H, Weng C, Fasel D, Knollmann BC, Wang TJ, Denny JC, Ellinor PT, Roden DM, Mosley JD. Association of Thyroid Function Genetic Predictors With Atrial Fibrillation: A Phenome-Wide Association Study and Inverse-Variance Weighted Average Meta-analysis. JAMA Cardiol 2019;**4**(2):136-43 doi: 10.1001/jamacardio.2018.4615[published Online First: Epub Date]|.
58. Hardwick SA, Deveson IW, Mercer TR. Reference standards for next-generation sequencing. Nat Rev Genet 2017;**18**(8):473-84 doi: 10.1038/nrg.2017.44[published Online First: Epub Date]|.
59. Liao KP, Sparks JA, Hejblum BP, Kuo IH, Cui J, Lahey LJ, Cagan A, Gainer VS, Liu W, Cai TT, Sokolove J, Cai T. Phenome-Wide Association Study of Autoantibodies to Citrullinated and Noncitrullinated Epitopes in Rheumatoid Arthritis. Arthritis Rheumatol 2017;**69**(4):742-49 doi: 10.1002/art.39974[published Online First: Epub Date]|.
60. Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, Richardson D, Poole C. Illustrating bias due to conditioning on a collider. Int J Epidemiol 2010;39(2):417-20 doi: 10.1093/ije/dyp334[published Online First: Epub Date]|.
61. Verma A, Ritchie MD. Current Scope and Challenges in Phenome-Wide Association Studies. Curr Epidemiol Rep 2017;**4**(4):321-29 doi: 10.1007/s40471-017-0127-7[published Online First: Epub Date]|.
62. Karnes JH, Bastarache L, Shaffer CM, Gaudieri S, Xu Y, Glazer AM, Mosley JD, Zhao S, Raychaudhuri S, Mallal S, Ye Z, Mayer JG, Brilliant MH, Hebbring SJ, Roden DM, Phillips EJ, Denny JC. Phenome-wide scanning identifies multiple diseases and disease severity phenotypes associated with HLA variants. Sci Transl Med 2017;**9**(389) doi: 10.1126/scitranslmed.aai8708[published Online First: Epub Date]|.
63. Sinnott JA, Cai F, Yu S, Hejblum BP, Hong C, Kohane IS, Liao KP. PheProb: probabilistic phenotyping using diagnosis codes to improve power for genetic association studies. J Am Med Inform Assoc 2018;**25**(10):1359-65 doi: 10.1093/jamia/ocy056[published Online First: Epub Date]|.
64. Cordova-Palomera A, Tcheandjieu C, Fries JA, Varma P, Chen VS, Fiterau M, Xiao K, Tejeda H, Keavney BD, Cordell HJ, Tanigawa Y, Venkataraman G, Rivas MA, Ré C, Ashley E, Priest JR. Cardiac Imaging of Aortic Valve Area From 34 287 UK Biobank Participants Reveals Novel Genetic Associations and Shared Genetic Comorbidity With Multiple Disease Phenotypes. Circ Genom Precis Med 2020;**13**(6):e003014 doi: 10.1161/CIRCGEN.120.003014[published Online First: Epub Date]|.
65. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat Rev Genet 2018;**19**(8):491-504 doi: 10.1038/s41576-018-0016-z[published Online First: Epub Date]|.
66. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods 2010;**7**(4):248-9 doi: 10.1038/nmeth0410-248[published Online First: Epub Date]|.
67. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 2009;**4**(7):1073-81 doi: 10.1038/nprot.2009.86[published Online First: Epub Date]|.
68. Consortium EP. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 2004;**306**(5696):636-

40 doi: 10.1126/science.1105136[published Online First: Epub Date]|.

69. Consortium GT. The Genotype-Tissue Expression (GTEx) project. Nat Genet 2013;**45**(6):580-5 doi: 10.1038/ng.2653[published Online First: Epub Date]|.
70. The Gene Ontology C. Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res 2017;**45**(D1):D331-D38 doi: 10.1093/nar/gkw1108[published Online First: Epub Date]|.
71. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res 2004;**32**(Database issue):D277-80 doi: 10.1093/nar/gkh063[published Online First: Epub Date]|.
72. Pulley JM, Shirey-Rice JK, Lavieri RR, Jerome RN, Zaleski NM, Aronoff DM, Bastarache L, Niu X, Holroyd KJ, Roden DM, Skaar EP, Niswender CM, Marnett LJ, Lindsley CW, Ekstrom LB, Bentley AR, Bernard GR, Hong CC, Denny JC. Accelerating Precision Drug Development and Drug Repurposing by Leveraging Human Genetics. Assay Drug Dev Technol 2017;**15**(3):113-19 doi: 10.1089/adt.2016.772[published Online First: Epub Date]|.
73. Rastegar-Mojarad M, Ye Z, Kolesar JM, Hebbring SJ, Lin SM. Opportunities for drug repositioning from phenome-wide association studies. Nat Biotechnol 2015;**33**(4):342-5 doi: 10.1038/nbt.3183[published Online First: Epub Date]|.
74. Millwood IY, Bennett DA, Walters RG, Clarke R, Waterworth D, Johnson T, Chen Y, Yang L, Guo Y, Bian Z, Hacker A, Yeo A, Parish S, Hill MR, Chissoe S, Peto R, Cardon L, Collins R, Li L, Chen Z, China Kadoorie Biobank Collaborative G. A phenome-wide association study of a lipoprotein-associated phospholipase A2 loss-of-function variant in 90 000 Chinese adults. Int J Epidemiol 2016;**45**(5):1588-99 doi: 10.1093/ije/dyw087[published Online First: Epub Date]|.
75. Jerome RN, Joly MM, Kennedy N, Shirey-Rice JK, Roden DM, Bernard GR, Holroyd KJ, Denny JC, Pulley JM. Leveraging Human Genetics to Identify Safety Signals Prior to Drug Marketing Approval and Clinical Use. Drug Saf 2020;**43**(6):567-82 doi: 10.1007/s40264-020-00915-6[published Online First: Epub Date]|.
76. Diogo D, Tian C, Franklin CS, Alanne-Kinnunen M, March M, Spencer CCA, Vangjeli C, Weale ME, Mattsson H, Kilpelainen E, Sleiman PMA, Reilly DF, McElwee J, Maranville JC, Chatterjee AK, Bhandari A, Nguyen KH, Estrada K, Reeve MP, Hutz J, Bing N, John S, MacArthur DG, Salomaa V, Ripatti S, Hakonarson H, Daly MJ, Palotie A, Hinds DA, Donnelly P, Fox CS, Day-Williams AG, Plenge RM, Runz H. Phenome-wide association studies across large population cohorts support drug target validation. Nat Commun 2018;**9**(1):4285 doi: 10.1038/s41467-018-06540-3[published Online First: Epub Date]|.
77. Gill D, Georgakis MK, Koskeridis F, Jiang L, Feng Q, Wei WQ, Theodoratou E, Elliott P, Denny JC, Malik R, Evangelou E, Dehghan A, Dichgans M, Tzoulaki I. Use of Genetic Variants Related to Antihypertensive Drugs to Inform on Efficacy and Side Effects. Circulation 2019;**140**(4):270-79 doi: 10.1161/CIRCULATIONAHA.118.038814[published Online First: Epub Date]|.
78. Rao AS, Lindholm D, Rivas MA, Knowles JW, Montgomery SB, Ingelsson E. Large-Scale Phenome-Wide Association Study of PCSK9 Variants Demonstrates Protection Against Ischemic Stroke. Circ Genom Precis Med 2018;**11**(7):e002162 doi: 10.1161/CIRCGEN.118.002162[published Online First: Epub Date]|.
79. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, Pacheco JA, Tromp G, Pathak J, Carrell DS, Ellis SB, Lingren T, Thompson WK, Savova G, Haines J, Roden DM, Harris PA, Denny

JC. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc 2016;**23**(6):1046-52 doi: 10.1093/jamia/ocv202[published Online First: Epub Date]|.