Deep-PheWAS: a pipeline for phenotype generation and association analysis for phenome-wide association studies.

Packer R¹,², Williams AT¹, Hennah W³⁵,⁶, Eisenberg MT², Fawcett KA¹, Pearson W¹, Guyatt A¹, Edris A³¹, Hollα EI², Rao BS⁷, Bratty JR⁷, Wain LV¹,², Dudbridge F¹, Tobin MD¹,²

1. Department of Health Sciences, University of Leicester, Leicester, UK.
2. Department of Genetics and Genome Biology, University of Leicester, Leicester, UK
3. Department of Bioanalysis, Faculty of Pharmaceutical Sciences, Ghent University, Ghent, Belgium
4. Orion Pharma, Espoo, Finland
5. Neuroscience Center, HiLIFE, University of Helsinki, Helsinki, Finland
6. Institute for Molecular Medicine FIMM, HiLIFE, University of Helsinki, Helsinki, Finland
7. Orion Pharma, Nottingham, UK
8. Leicester National Institute of Health and Care Research Biomedical Research Centre, Glenfield Hospital, Leicester, UK

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Abstract

Summary: Deep-PheWAS is a platform for phenome wide association studies that creates clinically-curated composite phenotypes, and integrates quantitative phenotypes from primary care data, longitudinal trajectories of quantitative measures, disease progression, and drug response phenotypes. Tools are provided for efficient analysis of association with any genetic input, under any genetic model, with optional sex-stratified analysis, and for developing novel phenotypes.

Availability and Implementation: The Deep-PheWAS pipeline is freely available under GNU general public licence v3.0 at https://github.com/Richard-Packer/Deep-PheWAS.

Contact: richard.packer@leicester.ac.uk

Supplementary information: Supplementary methods and results are available at Bioinformatics online.
Introduction

Phenome-wide association studies (PheWASs) can be used to better understand the pleiotropic effects of genetic variants (Tyler et al., 2016) and to inform drug development through target identification, target validation and use of variants that mimic drug effects to assess likely drug efficacy, safety and drug repurposing opportunities (Gill et al., 2019; Diogo et al., 2018; Khosravi et al., 2019). PheWASs comprise two stages – phenotype generation and statistical association tests. There have been two widely applicable methods for phenotype generation: PHESANT (Millard et al., 2018) and PheWAS-R (Carroll et al., 2014). PHESANT creates phenotypes by extracting study-specific questionnaire and measurement data alongside linked hospital records in UK Biobank. PheWAS-R combines related international classification of disease version 9 and 10 (ICD-9/ICD-10) codes into clinically relevant groups termed Phecodes. Both tools provide regression analysis for per-variant PheWAS for generated phenotypes using generalised linear models in R and produce Manhattan plots. Online PheWAS resources such as Open Targets Genetics (Ghoussaini et al., 2021) do not perform new statistical tests. Instead, they are repositories for existing results from phenotypes generated by one of the two above tools or by individual genome wide association studies (GWAS).

These tools are useful, but have several key gaps:

1) The phenotypes generated rely on a single data field or coding ontology and do not take advantage of all available data, such as primary care data;
2) Existing approaches do not provide tools for developing new phenotypes;
3) For running per-variant PheWAS, running each regression model in R is computationally inefficient and can result in inflated type I error for low-frequency variants with a case-control imbalance (Ma et al., 2013);
4) Online resources such as Open Targets Genetics have limited flexibility. For example, they accept only single nucleotide polymorphisms (SNPs), and retrieve results only for genetic models tested. The user cannot specify when to use new data fields or updates to existing data fields (for example, updated health records), and the user cannot specify their preferred statistical approach and outputs, such as false discovery rate (FDR).

The platform we have developed, Deep-PheWAS, addresses both phenotype generation and efficient association testing while incorporating the following developments that are not yet available in any current platform or online resource:

(i) Clinically-curated composite phenotypes for selected health conditions that integrate different data types (including primary and secondary care data) to study phenotypes not well captured by current classification trees;
(ii) Integration of quantitative phenotypes from primary care data, such as pathology records and clinical measures;
(iii) Integration of disease progression phenotypes, longitudinal trajectories of quantitative measures and drug response measures;
(iv) Clinically-curated phenotype selection for traits that are extremely highly correlated;
(v) Flexible tests of additive, dominant, recessive and genotypic models;
(vi) Inclusion of complex structural variants, such as multi-allelic CNVs;
(vii) Ability to test genetic risk scores;
(viii) Creation of phenotypes in sex-specific strata to run a sex-stratified PheWAS;
(ix) Providing tools for generating novel phenotypes using a simple phenotype mapping process.
We provide an overview of features of Deep-PheWAS compared to those of previously published platforms in Supplementary Table 1 and illustrate these below.

**Application of Deep-PheWAS to UK Biobank**

**Analysis of quantitative phenotypes**

Our platform can be applied to quantitative phenotypes derived from numerous data sources, including primary care data. For example, we created a phenotype using recorded levels of blood sodium in primary care records that is not yet included in any PheWAS platform. We applied Deep-PheWAS to rs7193778 (nearest genes NFAT5 and TERF2), previously associated with urate levels (Köttgen et al., 2013). Our PheWAS shows various associations which are currently not documented in GWAS Catalog, most strongly with blood sodium levels (Supplementary Figure 1, Supplementary Table 2).

**Highly correlated traits**

We applied our Deep-PheWAS approach to rs2912062 (nearest genes ANGPT2 and AGPAT5), shown to be associated with carotid intima-media thickness (IMT) (Strawbridge et al., 2020), a phenotype not currently available in any PheWAS platform. By selecting a single representative measure taken from many individual measurements Deep-PheWAS can collapse highly correlated quantitative traits into single measures (in this case carotid IMT maximum and carotid IMT mean), reducing redundancy and improving power. We recapitulated known GWAS findings (Supplementary Figure 2, Supplementary Table 3).

**Flexibility with choice of genetic model**

To demonstrate the flexibility of Deep-PheWAS we assessed the SERPINA1 Z allele (rs28929474(T)) under additive and recessive genetic models. Patterns of association differed between models, including opposite estimated effect directions for this variant on forced expiratory volume in 1 second (FEV₁) and forced vital capacity (FVC), consistent with previous findings (Fawcett, Song, et al., 2021) (Supplementary Figures 3 and 4, Supplementary Tables 4 and 5).

**Association tests for complex structural variation**

Human genomic variation includes variants which have more categories than SNPs. For example, diploid human copy number of CCL3L1 ranges from 0 to 8 in UK Biobank participants (Fawcett, Demidov, et al., 2021) (Supplementary Figure 5). In such situations, association testing may be based on the measured copy number or on user-specified collapsed categories, requiring a flexible platform. We used Deep-PheWAS to test association with CCL3L1 copy number (coded 0-8) under a linear additive model); no associations reached an FDR threshold of 1% (Supplementary Figure 6, the top 5 associations are shown in Supplementary Table 6), this recapitulates findings from earlier studies (Field et al., 2009; Urban et al., 2009; Carpenter et al., 2011; Adewoye et al., 2018).

**Genetic risk scores, composite, and disease-progression phenotypes**

Genetic risk scores (GRS) aggregate multiple SNPs, providing improved power for studying phenotypic associations, but cannot be specified in online PheWAS platforms. We performed a PheWAS using a 279-variant GRS for FEV₁/FVC (Shrine N, et al. 2019), which showed association (FDR<1%) with 47 traits including increased risk of clinical COPD and clinical asthma with a higher score of FEV₁/FVC reducing alleles (Supplementary Figure 7, Supplementary Table 7). Furthermore, the composite phenotypes generated by the Deep-PheWAS platform (for example P2020 Asthma and P2054 COPD) were consistently more strongly associated with the GRS for FEV₁/FVC than the...
relevant Phecodes alone. We also show significant association with the novel disease-progression phenotypes: acute exacerbations of COPD (AECOPD) and age-of-onset of COPD both of which are unavailable in existing PheWAS resources and have published GWAS results.

Sex-stratified PheWAS

We applied Deep-PheWAS to rs12777332 (nearest genes CASP7 and NRAP) and rs7697189 (nearest gene HHIP). We replicate association for risk of cataract in females only (Supplementary Figure 8, Supplementary Table 8) for rs12777332 (Choquet et al., 2021), and for increased effect on FEV₁ in males compared to females (Supplementary Figures 9 and 10, Supplementary Table 9) for rs7697189 (Fawcett, Obeidat, et al., 2021).

Implementation

Deep-PheWAS is a suite of R scripts that can be run on high-performance computing clusters and requires R 4.1.0 and PLINK 2.0. Deep-PheWAS is optimised for UK Biobank data and is expected to be interoperable with the UK Biobank Research Analysis Platform but can also be applied to other data sources, see user guide.

Availability

The Deep-PheWAS pipeline is freely available under GNU general public licence v3.0 at https://github.com/Richard-Packer/Deep-PheWAS.

Conclusion

Here, we present Deep-PheWAS, a suite of open-source R scripts that facilitates phenome-wide association studies while addressing several limitations of existing approaches. This includes the ability to analyse a broader range of phenotypes derived from large-scale electronic healthcare records, more informative composite phenotypes, greater flexibility in the type of genetic variation that can be studied and assessing associations with genetic risk scores.

Acknowledgements

This research has been conducted using UK Biobank Resource under Application Number 43027. This research used the SPECTRE High Performance Computing Facility at the University of Leicester.

Funding

This work is supported by:

Orion Pharma funded a research collaboration with the University of Leicester, which supported the development of Deep-PheWAS.

ALG: Wellcome Trust Institutional Strategic Support Fund [204801/Z/16/Z], BHF Accelerator Award [AA/18/3/34220]. LVW: GSK/Asthma + Lung UK Chair in Respiratory Research (C17-1). MDT: Wellcome Trust Investigator Award [WT202849/Z/16/Z]. MDT holds a National Institute for Health and Care Research Senior Investigator Award. The respiratory research was partially supported by BREATHE — The Health Data Research Hub for Respiratory Health [MC_PC_19004]. The research was partially supported by the National Institute for Health and Care Research Leicester Biomedical
Research Centre; views expressed are those of the author(s) and not necessarily those of the NHS, the National Institute for Health and Care Research or the Department of Health. This research was funded in part by the Wellcome Trust. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Declaration of Interests
MDT, RP, ATW, FD and LVW received funding from Orion Pharma within the scope of the submitted work, and from GSK for collaborative research projects outside of the submitted work. WH, BSR and RB are salaried employees of Orion Pharma within the scope of the submitted work. LVW has performed consultancy for Galapagos and had travel paid for by Genetech outside the scope of this work.

Data availability
The data underlying this article is available from the UK Biobank to all approved researchers https://www.ukbiobank.ac.uk/. Data was accessed under approved application 43027.

References


