**ABSTRACT**

Summary: Phenome-wide association studies (PheWAS) have been used to replicate known genetic associations and discover new phenotype associations for genetic variants. This PheWAS implementation allows users to translate ICD-9 codes to PheWAS case and control groups, perform analyses using these and/or other phenotypes with covariate adjustments and plot the results. We demonstrate the methods by replicating a PheWAS on rs3135388 (near HLA-DRB, associated with multiple sclerosis) and performing a novel PheWAS using an individual’s maximum white blood cell count (WBC) as a continuous measure. Our results for rs3135388 replicate known associations with more significant results than the original study on the same dataset. Our PheWAS of WBC found expected results, including associations with infections, myeloproliferative diseases and associated conditions, such as anemia. These results demonstrate the performance of the improved classification scheme and the flexibility of PheWAS encapsulated in this package.

Availability and implementation: This R package is freely available under the Gnu Public License (GPL-3) from http://phewascatalog.org. It is implemented in native R and is platform independent.

Contact: phewas@vanderbilt.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on November 12, 2013; revised on April 3, 2014; accepted on April 9, 2014

2 METHODS

2.1 Data input

Data to perform a PheWAS using this package can include, for a population, the following: demographic data, International Classification of Disease codes diagnostic code data, the independent variables (e.g. genotype or laboratory data) and any other covariates, such as principal components generated to adjust for genetic ancestry. Users can pass into the phewas method any data types R supports for regression.

Although the original PheWAS study used genetic data as predictors, studies using phenotypes as predictors are also feasible in this framework.

2.2 Mapping phenotypes

Investigators can perform a PheWAS using ICD-9 codes or ‘PheWAS codes’, which represent ~1600 hierarchical phenotypes formed from grouped ICD-9 codes. Each PheWAS phenotype also includes an optional set of exclusion phenotypes for similar diagnoses to more accurately identify true controls. This step prevents patients with common ‘rule out’ codes or similar diseases from being marked as a control during the statistical analysis (e.g. a patient with an unknown arrhythmia cannot serve as a control for atrial fibrillation). As requiring multiple codes occurring on different days for a given diagnosis improves phenotype precision, users can specify count thresholds required to establish a patient as a ‘case’ for a given phenotype or simply use the code count in the regression model. This threshold is often set to a minimum of two unique code days (Denny et al., 2013).

2.3 Statistical analysis

Users can choose among different statistical tests when performing a PheWAS, including adjusted and unadjusted models. The default analyses are linear or logistic regression. $\chi^2$ and $t$-tests are available for fast unadjusted tests. $P$-values, betas, case and control counts and odds ratios

*To whom correspondence should be addressed.
We found an association between multiple sclerosis and rs3135388 with an OR of 2.56 and a *P*-value of $2.8 \times 10^{-6}$. This is an improvement over the previous analysis that included 733 phenotypes and used a $\chi^2$ test. The second example applied linear regression to PheWAS codes adjusted for age and gender to predict maximum white blood cell count (WBC) in the same population.

### 3 RESULTS AND DISCUSSION

We found an association between multiple sclerosis and rs3135388 with an OR of 2.56 and a *P*-value of $1.4 \times 10^{-7}$ (Fig. 1). The improved methods, including a covariate-adjusted analysis and revised phenotypes, yielded an OR more consistent with the largest published study on this association, which reported an OR of 2.75 (De Jager et al., 2009). Maximum WBC was associated with infections, leukemias and other expected conditions (Fig. 2).

These analyses were performed from ICD-9 codes and demographic data using the createPhewsTable, phews, and phewsManhattan methods of the R PheWAS package. The top lines represent Bonferroni significance, and differing output options were used to showcase functionality. Supplementary Tables S1 and S2 include the top 25 hits by *P*-value for each analysis.

As more investigators leverage EMR data for clinical and genomic analyses, available validated methods will become more valuable. These methods should permit easier adoption of EMR-based PheWAS by more researchers. As shown in the WBC analysis, the PheWAS methodology can also be applied to non-genetic data, providing new avenues of investigation for PheWAS.

**Funding:** The project was supported by National Library of Medicine (5T15LM007450 and R01 LM010685).

**Conflict of Interest:** none declared.

### REFERENCES


