Phenome-wide association studies on cardiovascular health and fatty acids considering phenotype quality control practices for epidemiological data[*]

Kristin Passero and Xi He
Huck Institutes of the Life Sciences, The Pennsylvania State University
University Park, PA 16802, USA
Email: kxp642@psu.edu, xuh142@psu.edu

Jiayan Zhou
Department of Veterinary and Biomedical Sciences, The Pennsylvania State University,
University Park, PA 16802, USA
Email: jpz5091@psu.edu

Bertram Mueller-Myhsok
Statistical Genetics, Max Planck Institute of Psychiatry,
80804 Munich, Germany
Munich Cluster of Systems Biology, SyNergy,
81377 Munich, Germany
Institute of Translational Medicine, University of Liverpool,
Liverpool L69 3BX, United Kingdom
Email: bmm@psych.mpg.de

Marcus E. Kleber
Vth Department of Medicine, Medical Faculty Mannheim, Heidelberg University,
Mannheim, Germany
Email: marcus.kleber@medma.uni-heidelberg.de

Winfried Maerz
Vth Department of Medicine, Medical Faculty Mannheim, Heidelberg University,
Mannheim, Germany
Synlab Holding Deutschland GmbH, SYNLAB Academy,
Mannheim and Augsburg, Germany
Medical University of Graz, Clinical Institute of Medical and Chemical Laboratory Diagnostics,
Graz, Austria
Email: winfried.maerz@synlab.com

Molly A. Hall
Department of Veterinary and Biomedical Sciences, The Pennsylvania State University,
University Park, PA 16802, USA
Email: mah546@psu.edu

[*] This work is supported by the USDA National Institute of Food and Agriculture and Hatch Appropriations under Project #PEN04275 and Accession #1018544

Phenome-wide association studies (PheWAS) allow agnostic investigation of common genetic variants in relation to a variety of phenotypes but preserving the power of PheWAS requires careful phenotypic quality control (QC) procedures. While QC of genetic data is well-defined, no established QC practices exist for multi-phenotypic data. Manually imposing sample size restrictions, identifying variable types/distributions, and locating problems such as missing data or outliers is arduous in large, multivariate datasets. In this paper, we perform two PheWAS on epidemiological data and, utilizing the novel software CLARITE (CLeaning to Analysis: Reproducibility-based Interface for Traits and Exposures), showcase a transparent and replicable phenome QC pipeline which we believe is a necessity for the field. Using data from the Ludwigshafen Risk and Cardiovascular (LURIC) Health Study we ran two PheWAS, one on cardiac-related diseases and the other on polyunsaturated fatty acids levels. These phenotypes underwent a stringent quality control screen and were regressed on a genome-wide sample of single nucleotide polymorphisms (SNPs). Seven SNPs were significant in association with dihomo-γ-linolenic acid, of which five were within fatty acid desaturases *FADS1* and *FADS2*. PheWAS is a useful tool to elucidate the genetic architecture of complex disease phenotypes within a single experimental framework. However, to reduce computational and multiple-comparisons burden, careful assessment of phenotype quality and removal of low-quality data is prudent. Herein we perform two PheWAS while applying a detailed phenotype QC process, for which we provide a replicable pipeline that is modifiable for application to other large datasets with heterogenous phenotypes. As investigation of complex traits continues beyond traditional genome wide association studies (GWAS), such QC considerations and tools such as CLARITE are crucial to the in the analysis of non-genetic big data such as clinical measurements, lifestyle habits, and polygenic traits.

Keywords: PheWAS; Phenotype; Quality control; Precision medicine.

1. Introduction

Phenome-wide association studies (PheWAS) seek to identify factors significantly associated with multiple human-health related traits. Unlike genome-wide association studies (GWAS), which analyze a single phenotype of interest with respect to genome-wide genetic variation, PheWAS leverage that genetic variation against complex, phenome-wide information. The consideration of multiple phenotypes allows PheWAS to assess the interconnectedness of health conditions, making them useful tools for follow-up investigation of risk loci identified in GWAS[1,2,3]. PheWAS can identify statistical pleiotropy or potentially elucidate gene networks as part of disease etiology[3,4]. Like GWAS, they suffer from a great multiple testing burden and incompatibility with rare variant analysis[3,5].

An additional difficulty of analyzing phenome-wide datasets is the breadth and heterogeneity of the phenotypes, which may require different quality control (QC) considerations. Due to the popularity of leveraging electronic medical records (EMR) to acquire phenotypic information, phenome QC has focused on properly condensing EMR's International Classification of Diseases (ICD) codes into distinct phenotypes[2,6,7]. Successful PheWAS using ICD-classification algorithms are well-documented[1,2,6,7,8,9]. Yet phenotypic classification is but one aspect of phenotype QC; less focus has been given to the simple but easily applicable idea of data cleaning. Evaluating sample size, data missingness, and identifying documentation errors are hurdles that efficient phenotype QC must overcome to define a high-quality phenome. Post-phenotype-classification QC excludes phenotypes with low variability or sample size that may needlessly increase the multiple test burden. Furthermore, whereas ICD codes enable neat definition of case-control phenotypes, data from epidemiologic studies may document non-binary phenotypes. Such data must be perused, and phenotypes identified by variable type to correctly apply analytic methods, such as logistic or linear regression. However, given the large size of phenomic datasets, manual inspection of data is inefficient. Furthermore, replicable and reproducible PheWAS rely on the transparency of the QC workflow. Researchers benefit from data-cleaning tools with broad filtering and classification capabilities for simultaneous QC of phenotypes, but which also rely on sufficient user input to establish accountability for and leave a working record of the QC decisions made. Here we document QC of phenotypic data for categorical and quantitative traits, using data from the Ludwigshafen Risk and Cardiovascular (LURIC) Health Study and implementing QC in the CLARITE software (Lucas, Palmiero et al., 2019, accepted; documentation and code at https://github.com/HallLab). We performed two PheWAS respectively utilizing binary or continuous phenotypes related to cardiac health and coronary artery disease. We provide detailed QC pipelines for both (Fig. 1; Suppl. M1. Supplementary material is accessible at https://drive.google.com/file/d/1a8twpSL9Hvk95gsx6piBuo1ZuHb60wPv/view?usp=sharing). Phenotypes included disease diagnoses (e.g., coronary artery disease, peripheral vascular disease), associated risk factors (e.g., diabetes, lipid serum metabolites, hypertension), and follow-up mortality (Table S1). We ran an analysis while applying a proposed phenotype quality control pipeline, implemented in CLARITE which is designed to facilitate reproducible quality control workflows. We tested over 500,000 SNPs available in the LURIC data for association with case-control and quantitative phenotypes associated with development of cardiovascular disease and replicated several known associations of genetic variants with dihomo-γ-linolenic acid. We demonstrate a QC pipeline for raw phenotypic data (Fig. 1; Suppl. M1), offer suggestions for best practices of phenome QC, and recommend CLARITE as a means of its efficient and transparent enaction. Refining phenotype QC is an easily adoptable practice to improve quality of PheWAS, thereby decreasing risk of spurious associations; CLARITE conveniently provides the tools for QC in a single package, facilitating QC for PheWAS utilizing high-dimensional data. It further accommodates investigations with heterogeneous or multiple sources of health data (exposures, biomarkers, clinical features, etc.) and facilitates the implementation of data preprocessing practices in PheWAS designed to screen multiple disease traits or environmental risk factors.

2. Methods

2.1. *Ludwigshafen risk and cardiovascular health study*

The Ludwigshafen Risk and Cardiovascular (LURIC) Health Study, is a prospective cohort evaluating genetic and pharmacological risk factors of cardiovascular diseases and other associated phenotypes[10]. Beginning in 1997, the study focused on the predictors of coronary artery disease (CAD), myocardial infarctions (MI), Type II diabetes (T2D), and hypertension, given that they are prevalent in Western culture[10]. Participants were of white European/German ancestry and required a coronary angiogram, either previously obtained or performed at the Ludwigshafen Heart Centre prior to participation, to appropriately classify CAD[10]. Additionally, no participant had been diagnosed with a non-cardiac disease, was recovering from surgery, nor had a history of cancer[10]. All participants gave informed consent. LURIC collected data from a standardized questionnaire and a physical examination[10]. A blood sample was taken for fatty acid measures and an oral glucose tolerance test was ordered for diagnosis of diabetes[10]. Erythrocyte fatty acid composition was measured with the HS-Omega-3 Index methodology described previously[11]. Fatty acids were described as percentages out of total fatty acids identified. Our data contained information from the 2010 follow-up, and consisted of 3,316 individuals (30% female, 70% male; Table S2) from ages 18 – 95. The case-control ratio varied by phenotype and filtering by available genotype information restricted the number of cases (Table S1). The LURIC study is currently ongoing.

2.2. *Packages and tools*

Phenotype QC was performed in RStudio v1.1.463 with R v3.5.2 using CLARITE's R package distribution (cleaning phenome data) alongside *dplyr* (initial variable extraction), *haven* (reading .sav files), and *moments* (skew calculations). CLARITE is available as an R package (https://github.com/HallLab/clarite) and a Python package which accommodates a command line interface (https://www.hall-lab.org/clarite-python/). A GUI version of CLARITE is in development. CLARITE was our preferred tool for phenotype QC since its functions are designed for preprocessing of multivariate data. It expedites QC by simultaneous screening and/or processing of phenotypes. Its functions may be tailored to meet QC criteria particular to the user's needs and its user-directed implementation affords easy documentation of the QC process (Suppl. M2 & M3). CLARITE's suite of *get**-functions (*get_binary*, *get_categorical*, *get_continuous*) assign variable type according to the number of distinct values available in the data, which allow the user to separate qualitative and quantitative variables for independent, variable-specific QC, such as a transforming a quantitative trait or assessing case-control ratios of binary phenotypes. Additional features include concurrently screening sample size minimums, identifying unique values, recoding missing values, and producing bar plots, histograms, or frequency tables across phenotypes. While CLARITE does support chi-square- or linear/logistic regression-based tests for association, it does not read the binary files (BED/BIM/FAM) in which the LURIC genotype data was stored though this capability is in development. Phenotype data was exported as a .txt file for downstream analysis in PLATO[12], a software which can implement PheWAS. Data cleaned in CLARITE may be exported in a variety of formats to meet the requirements of the analysis software by utilizing any R function designed to export data.frames (e.g. *write.csv, write.table*). Genotype QC was implemented using PLINK[13] again due to CLARITE's current limitations in reading binary files. Visualization of Manhattan plots and p-value quantile-quantile plots were implemented with *ggplot2* and *qqman* respectively.

2.3. *Genotype and phenotype quality control*

The LURIC genotype data initially contained 3,061 samples (2,172 male and 889 female) and 687,262 SNPs. Prior to genotype QC, using data from the LURIC trait file, we filtered samples to remove pediatric cases (< 18 years) and persons missing covariate information (age, waist-hip ratio, BMI, and sex). Genotype QC was performed in PLINK for the remaining samples, first imposing a sex concordance check followed by 99% variant and sample call rates. SNPs with minor allele frequency (MAF) < 0.05 were removed. We checked for sample relatedness and, from each related pair (kinship coefficient > 0.125), removed the sample with greater missingness. Finally, we checked that the 99% sample/variant call rates and 5% MAF criteria were maintained following exclusion of related samples. The cleaned genotype data contained 2,824 samples (2,007 male and 817 female) and 577,007 SNPs. We calculated principal components (PCs) from the remaining samples and exported the first 3 PCs as covariates in our analysis.

A detailed description of and scripts for our phenotype QC are available in the Supplementary Material (Suppl. M1, M2 & M3). For our cardiac PheWAS, 33 disease phenotypes were available. Using CLARITE, we standardized missing values and selected only phenotypes with a minimum sample size of 200. Since PLATO does not perform multinomial/ordinal logistic regression, we defined variable types and restricted QC to binary phenotypes (no continuous phenotypes were among these selected traits). We filtered our data to retain only samples who passed

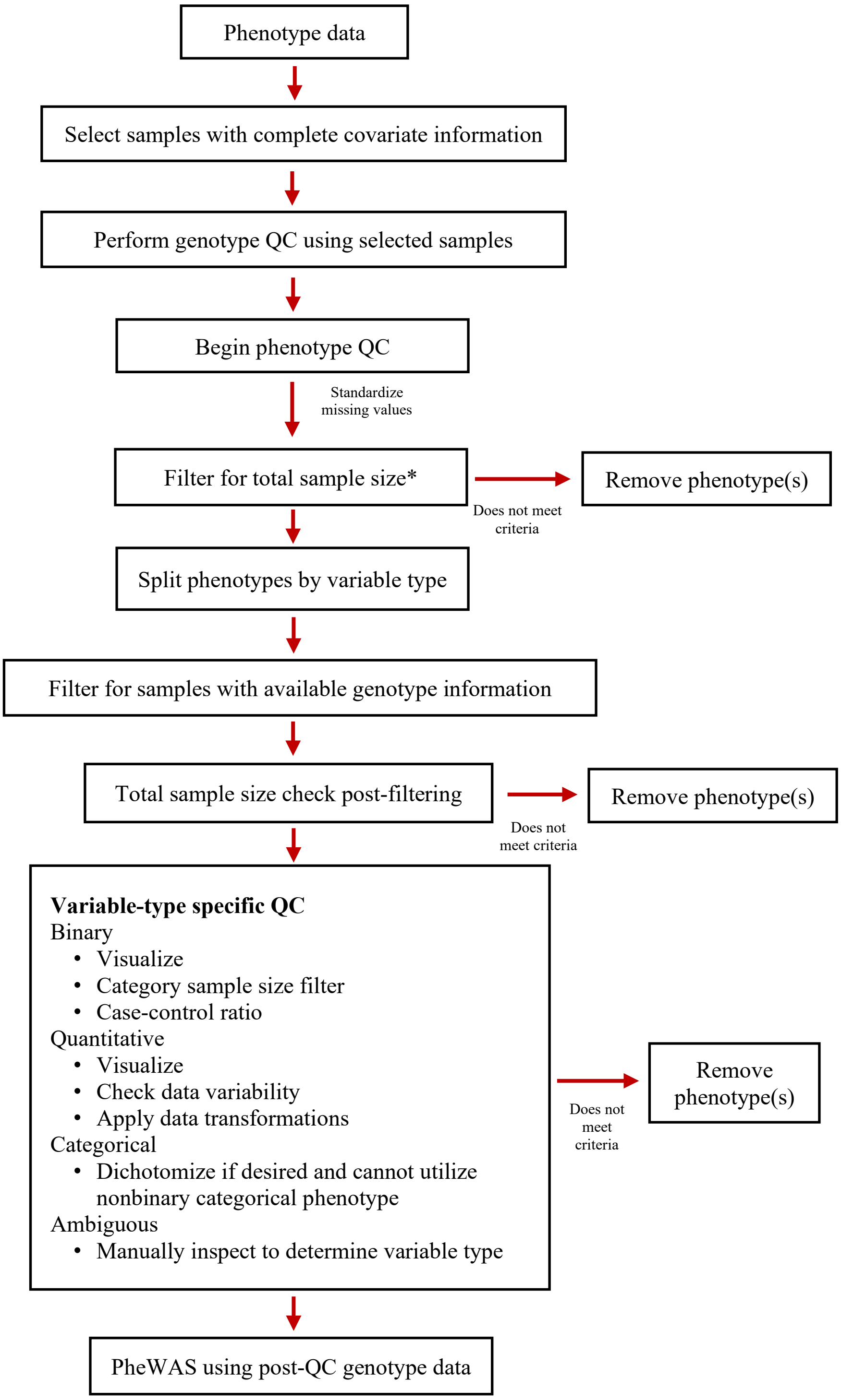

Phenotype data
Select samples with complete covariate information
Perform genotype QC using selected samples
Begin phenotype QC
Standardize missing values
Filter for total sample size*
Remove phenotype(s)
Does not meet criteria
Split phenotypes by variable type
Filter for samples with available genotype information
Total sample size check post-filtering
Remove phenotype(s)
Does not meet criteria
Variable-type specific QC
Binary
• Visualize
• Category sample size filter
• Case-control ratio
Quantitative
• Visualize
• Check data variability
• Apply data transformations
Categorical
• Dichotomize if desired and cannot utilize nonbinary categorical phenotype
Ambiguous
• Manually inspect to determine variable type
Remove phenotype(s)
Does not meet criteria
PheWAS using post-QC genotype data

Fig. 1. Suggested quality control pipeline for phenotype data. The same workflow was followed in our analyses. *The total sample size filter performed initially to minimize the number of phenotypes that needed to be sorted by variable type. This decision required an additional total sample seize check after filtering for samples who passed genotype QC.

genotype QC, which ensured the samples that remained contained no missing covariate information and has at least a 99% sample call rate. Finally, we required at least 150 cases and controls for each phenotype. Fifteen phenotypes remained for the cardiac PheWAS (Table S1). We considered 14 fatty acid phenotypes for our fatty acid PheWAS (Table S1). We filtered phenotypes to meet n = 200 minimum sample size and verified each were quantitative traits for linear regression in PLATO. Fatty acid phenotypes were filtered to keep only samples which passed genotype QC. Phenotypes were checked for skew; ten were highly skewed (skew > 0.5) and ln(1+x) transformed. Additionally, we ensured no phenotype contained greater than 90% zero values. All fatty acid phenotypes were included in the analysis. Final sample sizes varied due to differing missingness between phenotypes and are found in Table S1.

2.4. *Statistical analysis*

PLATO[12] used linear (fatty acid PheWAS) or logistic (cardiac PheWAS) regression to test the association of each SNP-phenotype combination, assuming an additive genetic model and adjusting for age, sex, BMI, waist-hip ratio, and the first 3 PCs. Fifteen cardiac disease or 14 fatty acid phenotypes were regressed against 577,007 SNPs. We investigated whether results met genome-wide significance ($p \leq 5\times10^{-8}$) and Bonferroni significance accounting for all tests run across both PheWAS ($p \leq 2.99\times10^{-9}$ corresponding to 0.05/[29×577,007]).

3. Results

3.1. *Cardiac PheWAS results*

While none of the SNPs tested in the cardiac PheWAS met Bonferroni nor genome-wide significance (Fig. S1; Table S3; Fig. S2), we investigated the top results. The SNP rs17701618 had the lowest p-value and was associated with stroke ($p = 2.40 \text{ x } 10^{-7}$, $\beta = -0.725$) and is 71kb downstream an intergenic long-noncoding RNA (lncRNA) *RP11-456A18.1*. Variant rs11701162 ($p = 3.28 \text{ x } 10^{-7}$, $\beta = -0.479$), associated with cardiomyopathy is an intronic variant of lncRNA *LINC01671*. SNP rs285584 was associated with mortality in the ten-year follow-up ($p = 9.57 \text{ x } 10^{-7}$, $\beta = -0.407$) and causes a missense mutation in the *PDZ Domain Containing Ring Finger 4 gene* (*PDZRN4*). Two intronic variants within *paralemmin 2* (*PALM2*) associated with Type II diabetes mellitus were rs4978846 ($p = 1.66 \text{ x } 10^{-6}$, $\beta = 0.360$) and rs10512400 ($p= 1.93 \text{ x } 10^{-6}$, $\beta = 0.357$).

3.2. *Fatty acid PheWAS results*

Seven SNPs met the genome-wide significance threshold (Fig. S3; Table S4; Fig. S4) and Bonferroni threshold ($p \leq 2.99\times10^{-9}$) in the fatty acid PheWAS; all were associated with dihomo-γ-linolenic acid (DGLA). Five statistically significant SNPs are intronic variants in the fatty acid desaturase genes *FADS1* and *FADS2*. The two variants in association with DGLA demonstrating the lowest p-values, rs174548 ($p = 2.06 \text{ x } 10^{-18}$, $\beta = -0.035$) and rs174549 ($p = 2.57 \text{ x } 10^{-18}$, $\beta = -0.035$), as well as the variant rs174547 ($p = 9.95 \text{ x } 10^{-14}$, $\beta = -0.029$), were within *FADS1*. SNPs rs174577 ($p = 10.2 \text{ x } 10^{-14}$, $\beta = -0.030$) and rs174583 ($p = 4.42 \text{ x } 10^{-14}$, $\beta = -0.029$)) were in *FADS2*. Two more significant SNPs were outside of the *FADS* genes. The variant rs4246215 (p = 6.42 x 10-

[15], β = -0.03) is in the 3' untranslated region (UTR) of *flap structure specific endonuclease 1 (FEN1)*. Another significant variant, rs174534 ($p = 3.24 \times 10^{-11}$, β = -0.026), is an intronic variant of *myelin regulatory factor (MYRF)*. The *FADS2* intronic variant, rs174577, was also nearly genome-wide significant in association with arachidonic acid ($p = 8.92 \times 10^{-8}$, β = 0.231).

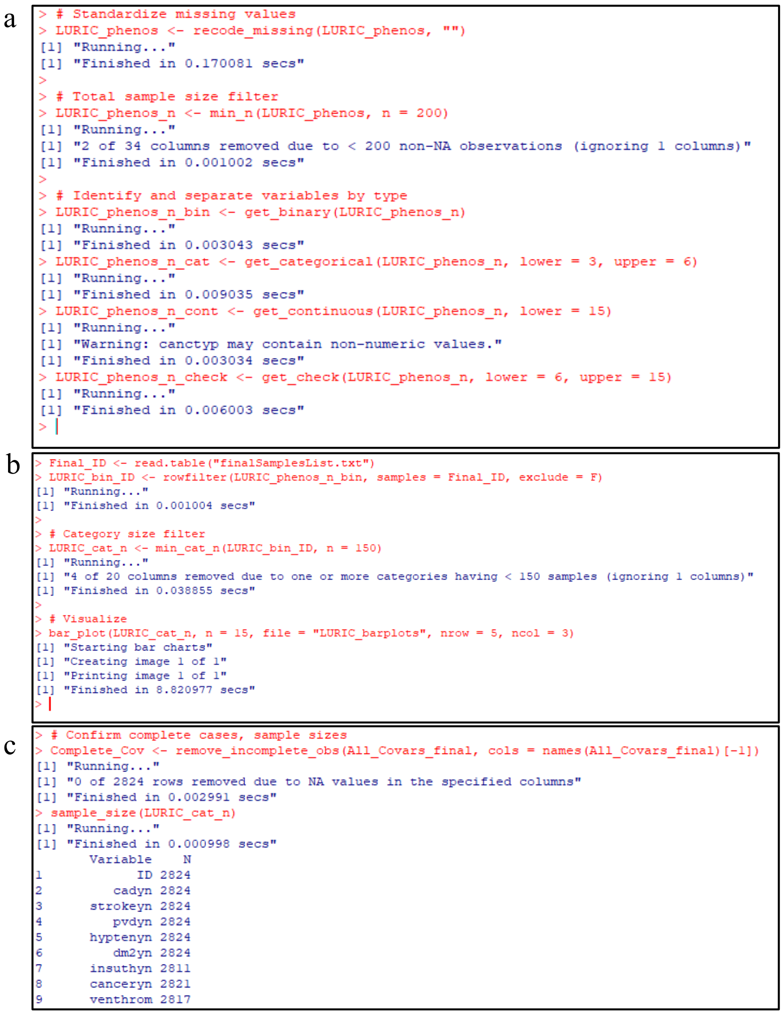
4. Discussion

In this study we performed two PheWAS, assessing associations between cardiovascular diseases/risk factors and genome-wide genetic variation while addressing rigorous yet replicable phenotype QC. The quality of PheWAS relies on the utilization of clean phenotype data. In part, ICD-code classification algorithms[6,1,7,9] address this and greatly facilitate PheWAS implementation with EMR data. However, not all PheWAS rely on EMR data, in which case the focus turns to designing an analysis with careful data cleaning procedures to obtain a high-quality phenome. Holistic and variable-type-specific QC checks are means of discarding substandard data from heterogenous, multivariate datasets, ensuring the preservation of high-quality phenotypes and reducing the multiple test burden by removing phenotypes which fail to meet quality standards. PheWAS using non-EMR data may particularly benefit from phenotype QC guidelines as the former may have nonuniform missing values, several variable types, or coding errors. We propose a workflow for phenotype QC (Fig. 1; Suppl. M1) that retains enough flexibility to ensure the researcher can choose parameters or preprocessing methods (e.g. sample size thresholds, transformations) pertinent to their data. We followed this pipeline on two PheWAS of our own, showcasing its applicability on both categorical and quantitative traits (Suppl. M2 & M3).

Our fatty acid PheWAS identified seven SNPs associated with dihomo-γ-linolenic acid (DGLA), many of which had previously documented associations with DGLA or FADS activity. A previous association found one such SNP, rs174548, was associated with percentage of DGLA out of total fatty acids levels[14]. Another study found an association between rs174548 and delta-6 desaturase (FADS2) activity, measured by dihomo-γ-linolenic acid: linolenic acid (DGLA:LA) ratio[15]. In our fatty acid PheWAS, rs174577 was significantly associated with DGLA and near-significantly associated with arachidonic acid (AA). Dorajoo et al. (2015) found the C allele of rs174577 was associated with increased AA in a Singaporean cohort. The direction of this association was reproduced in our results (Table S4). As such, this association has been found across ethnically diverse cohorts. While five of the seven significant SNPs were within the *FADS* locus, we found a significant association between DGLA and variant rs174534 in *MYRF*. Previous works have documented this variant's association with FADS1 activity (AA:DGLA ratio) in African Americans[16]. Additionally, while MYRF is foremost considered a transcription factor regulating myelination of the central nervous system[17], recent studies have implicated *de novo* mutations of *MYRF* in Scimitar syndrome and other congenital cardiac and urogenital defects[18,19,20].

The relationship between PUFAs, the *FADS* locus, and cardiovascular disease risk has been implicated in multiple association studies[21,22], but their functional relationship is complex. Chronic inflammation of the vasculature is characteristic of cardiovascular diseases and pro-inflammatory markers have been linked to incidence and more serious prognosis of CAD[23,24,25]. DGLA can be converted into anti-inflammatory metabolites PGE_1 and 15-HETrE[23,26]. However, FADS1 converts

DGLA into AA, which when further metabolized stimulates the proinflammatory response [26,27]. FADS1/2 further modulate the inflammatory response by converting α-linoleic acid (α-LA) into eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA), which themselves produce anti-inflammatory agents[27]. Altered rates of delta-5/FADS1 and delta-6/FADS2 desaturation, as measured by product-to-precursor PUFA ratios, have been found in cases of CAD[24] and T2D[28]. Multiple GWAS have found PUFA levels and desaturase activity are sensitive to FADS variants[14,15,16,24,26]. Further complicating the connection between PUFAs and CAD, humans obtain the linolenic acid precursors of DGLA, AA, EPA, and DHA through diet, which implies that a complex network of genetic and environmental factors governs cardiovascular disease. Accordingly, supplementation of "good" PUFAs has been considered a potential therapy for CAD, but studies document both successes and failures of PUFA supplementation to reduce cardiac events/mortality[29,23,27].



```
> # Standardize missing values
> LURIC_phenos <- recode_missing(LURIC_phenos, "")
[1] "Running..."
[1] "Finished in 0.170081 secs"
>
> # Total sample size filter
> LURIC_phenos_n <- min_n(LURIC_phenos, n = 200)
[1] "Running..."
[1] "2 of 34 columns removed due to < 200 non-NA observations (ignoring 1 columns)"
[1] "Finished in 0.001002 secs"
>
> # Identify and separate variables by type
> LURIC_phenos_n_bin <- get_binary(LURIC_phenos_n)
[1] "Running..."
[1] "Finished in 0.003043 secs"
> LURIC_phenos_n_cat <- get_categorical(LURIC_phenos_n, lower = 3, upper = 6)
[1] "Running..."
[1] "Finished in 0.009035 secs"
> LURIC_phenos_n_cont <- get_continuous(LURIC_phenos_n, lower = 15)
[1] "Running..."
[1] "Warning: canctyp may contain non-numeric values."
[1] "Finished in 0.003034 secs"
> LURIC_phenos_n_check <- get_check(LURIC_phenos_n, lower = 6, upper = 15)
[1] "Running..."
[1] "Finished in 0.006003 secs"
> |
```

```
> Final_ID <- read.table("finalSamplesList.txt")
> LURIC_bin_ID <- rowfilter(LURIC_phenos_n_bin, samples = Final_ID, exclude = F)
[1] "Running..."
[1] "Finished in 0.001004 secs"
>
> # Category size filter
> LURIC_cat_n <- min_cat_n(LURIC_bin_ID, n = 150)
[1] "Running..."
[1] "4 of 20 columns removed due to one or more categories having < 150 samples (ignoring 1 columns)"
[1] "Finished in 0.038855 secs"
>
> # Visualize
> bar_plot(LURIC_cat_n, n = 15, file = "LURIC_barplots", nrow = 5, ncol = 3)
[1] "Starting bar charts"
[1] "Creating image 1 of 1"
[1] "Printing image 1 of 1"
[1] "Finished in 8.820977 secs"
> |
```

```
> # Confirm complete cases, sample sizes
> Complete_Cov <- remove_incomplete_obs(All_Covars_final, cols = names(All_Covars_final)[-1])
[1] "Running..."
[1] "0 of 2824 rows removed due to NA values in the specified columns"
[1] "Finished in 0.002991 secs"
> sample_size(LURIC_cat_n)
[1] "Running..."
[1] "Finished in 0.000998 secs"
      Variable    N
1           ID 2824
2        cadyn 2824
3     strokeyn 2824
4        pvdyn 2824
5     hyptenyn 2824
6        dm2yn 2824
7     insuthyn 2811
8     canceryn 2821
9     venthrom 2817
```

Fig. 2. Example of phenome QC in CLARITE on binary phenotypes. Standardization of missing values, total sample size filter, and variable-type identification functions (a). Filtering of samples, category size filters, and data visualization (b). Removing incomplete cases and listing sample size (c).

Our cardiac PheWAS did not find genome-wide nor Bonferroni significant results which may be explained by differences in power when considering binary and quantitative responses. Quantitative variables like serum fatty acid levels preserve more phenotypic information as dichotomizing variables reduces statistical power and potentially hides variability within groups[30]. Defining cases and controls is necessary when lacking a clear quantitative trait to define a condition[31], but masks disease heterogeneity. CAD is diverse, with acute pathologies like myocardial infarction alongside chronic or asymptomatic etiologies[32]. Our sample had a majority of CAD cases. Phenotypes describing aspects of CAD, such as stroke, may have "controls" who were diagnosed with CAD but lacked that symptom. Thus, it may have been difficult to determine signal between affected and unaffected persons in the cardiac PheWAS due to overlap between case/control subgroups and potential for shared risk factors across controls and cases. Additional power may have been lost due to the case-only sample sizes available for our cardiac PheWAS. Apart from the phenotype "arrhythmia type", the overall sample sizes between the two PheWAS were similar (Table S1). However, in the cardiac PheWAS the number of cases ranged from about 170 to over 2,000. Phenotypes with smaller case-only sample sizes may be robust in a single analysis but not when accounting for the PheWAS multiple test burden. Furthermore, while our results corroborated previous research connecting the *FADS* locus' to PUFA levels, chronic inflammatory conditions, and cardiovascular diseases, the LURIC sample size was not sufficient to allocate a replication cohort for internal replication of significant associations. An additional limitation was the significance threshold imposed. Our PheWAS investigated joint-PheWAS Bonferroni-significant associations and results that met the genome-wide significance threshold. Multi-phenotypic analysis increases the multiple test burden provided each test is independent[33]. Since our phenotypes are related to a specific health trait, CAD, our analysis likely contains correlated phenotypes and a Bonferroni correction may be overly stringent[34].

We propose that rigorous phenomic QC be introduced to PheWAS methodology, a pipeline for which is displayed in Figure 1 and, with more detail, in the supplement (Suppl. M1). PheWAS already employ high-standards of genotype QC and often adopt filtering approaches to narrow down the variants of interest to a pathway or genomic region[4,6,7,35] thereby reducing the computational burden and multiple test penalty. Similar consideration of phenotypic data will ensure that phenotypes considered are of relative high-quality. While different studies necessitate different considerations, we propose certain QC criteria as best practices. Firstly, we recommend that studies identify and document a desired sample size minimum. Secondly, in heterogenous datasets, variable types should be identified and separated for individual processing. Thirdly, phenotype QC should only be performed on samples passing genotype QC. Finally, researchers should document thresholds used at and order of each QC step. In our analyses, we chose CLARITE to implement these practices due to its functions for concurrent cleaning of variables, its flexibility of function parameters, and its ease of maintaining documentation and tracking QC changes (Fig. 2; Suppl. M2 & M3).

CLARITE provides an interface for researchers to perform high-throughput data-cleaning prior to PheWAS. It is similar in function to the PHEnome Scan ANalysis Tool (PHESANT) [36] which performs phenome scans with UK Biobank data. PHESANT automatically classifies variables, performs linear, logistic, ordinal logistic, or multinomial logistic regression on available traits and described outcome(s), and visualizes results in QQ-plots or forest plots. The trade-off for the automated phenome scan afforded by PHESANT is its relative lack of user customization (it imposes certain sample size and distinct value restrictions for variable classification) and applicability only to UK BioBank data, as it requires UK Biobank information on data coding to identify missingness and ordinal variables. While CLARITE contains similar analytic capabilities, performing multivariate linear or logistic regression while additionally supporting analyses accounting for complex sampling design and survey weights when used in tandem with the R *survey* package[37], our focus is its usefulness as a tool for data cleaning. While more hands-on, CLARITE allows for modification to meet the researcher's needs and easily accounts for our suggested practices of phenotype QC.

While tools like CLARITE streamline QC and our proposed QC pipeline (Fig 1; Suppl. M1) can improve quality of analysis, they are limited by the relatively sparse investigation into the criteria of phenotype quality needed to improve PheWAS power and efficacy. Standards must be developed to determine a significance threshold analogous to the genome-wide threshold of GWAS and give better insight on what produces a sufficiently powered PheWAS. Much attention is given to filtering and QC of genotype data; phenotypes require the same considerations. In addition to genomic data, personal health records and data from epidemiological studies/surveys comprise a wealth of information on exposures or behaviors that contribute to individual health. A more rigorous perusal of phenotype quality through tools like CLARITE promotes use and eases implementation of analyses working to integrate this information to explore complex traits and multifaceted disease risk.

Acknowledgments

Genotyping of the LURIC study participants was supported by the 7th Framework Program AtheroRemo (grant agreement number 201668) of the European Union.

References

1. Denny JC, Crawford DC, Ritchie MD, et al. Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: Using electronic medical records for genome- and phenome-wide studies. *Am J Hum Genet*. 2011;89:529-542. doi:10.1016/j.ajhg.2011.09.008
2. Denny JC, Bastarache L, Ritchie MD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol*. 2013;31(12):1102-1110. doi:10.1038/nbt.2749
3. Hebbring SJ. The challenges, advantages and future of phenome-wide association studies. *Immunology*. 2014. doi:10.1111/imm.12195
4. Pendergrass SA, Brown-Gentry K, Dudek S, 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(1):e1003087. doi:10.1371/journal.pgen.1003087
5. Roden DM. Phenome-wide association studies: a new method for functional genomics in humans. *J Physiol*. 2017:4109-4115. doi:10.1113/JP273122
6. Denny JC, Ritchie MD, Basford MA, et al. PheWAS: Demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*. 2010;26(9):1205-1210. doi:10.1093/bioinformatics/btq126
7. Neuraz A, Chouchana L, Malamut G, et al. 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
8. Kathiresan S, Melander O, Anevski D, et al. Polymorphisms Associated with Cholesterol and Risk of Cardiovascular Events. *N Engl J Med*. 2008. doi:10.1056/nejmoa0706728
9. Verma A, Verma SS, Pendergrass SA, et al. EMERGE Phenome-Wide Association Study (PheWAS) identifies clinical associations and pleiotropy for stop-gain variants. *BMC Med*

Genomics. 2016;9(Suppl 1):32. doi:10.1186/s12920-016-0191-8
10. Winkelmann BR, März W, Boehm BO, et al. Rationale and design of the LURIC study - a resource for functional genomics, pharmacogenomics and long-term prognosis of cardiovascular disease. *Pharmacogenomics*. 2001;2(1s1):S1-S73. doi:10.1517/14622416.2.1.S1
11. Kleber ME, Delgado GE, Lorkowski S, März W, von Schacky C. *Trans* -fatty acids and mortality in patients referred for coronary angiography: the Ludwigshafen Risk and Cardiovascular Health Study. *Eur Heart J*. 2016;37(13):1072-1078. doi:10.1093/eurheartj/ehv446
12. Hall MA, Wallace J, Lucas A, et al. 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
13. Purcell S, Neale B, Todd-Brown K, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet*. 2007. doi:10.1086/519795
14. Mozaffarian D, Kabagambe EK, Johnson CO, et al. Genetic loci associated with circulating phospholipid trans fatty acids: A meta-analysis of genome-wide association studies from the CHARGE consortium. *Am J Clin Nutr*. 2015. doi:10.3945/ajcn.114.094557
15. Dorajoo R, Sun Y, Han Y, et al. A genome-wide association study of n-3 and n-6 plasma fatty acids in a Singaporean Chinese population. *Genes Nutr*. 2015. doi:10.1007/s12263-015-0502-2
16. Mathias RA, Sergeant S, Ruczinski I, et al. The impact of FADS genetic variants on ω6 polyunsaturated fatty acid metabolism in African Americans. *BMC Genet*. 2011. doi:10.1186/1471-2156-12-50
17. Bujalka H, Koenning M, Jackson S, et al. MYRF Is a Membrane-Associated Transcription Factor That Autoproteolytically Cleaves to Directly Activate Myelin Genes. *PLoS Biol*. 2013. doi:10.1371/journal.pbio.1001625
18. Chitayat D, Shannon P, Uster T, Nezarati MM, Schnur RE, Bhoj EJ. An Additional Individual with a De Novo Variant in Myelin Regulatory Factor (MYRF) with Cardiac and Urogenital Anomalies: Further Proof of Causality: Comments on the article by Pinz et al. (). *Am J Med Genet Part A*. 2018. doi:10.1002/ajmg.a.40360
19. Pinz H, Pyle LC, Li D, et al. De novo variants in Myelin regulatory factor (MYRF) as candidates of a new syndrome of cardiac and urogenital anomalies. *Am J Med Genet Part A*. 2018. doi:10.1002/ajmg.a.38620
20. Qi H, Yu L, Zhou X, et al. De novo variants in congenital diaphragmatic hernia identify MYRF as a new syndrome and reveal genetic overlaps with other developmental disorders. *PLoS Genet*. 2018. doi:10.1371/journal.pgen.1007822
21. Zhang JY, Kothapalli KSD, Brenna JT. Desaturase and elongase-limiting endogenous long-chain polyunsaturated fatty acid biosynthesis. *Curr Opin Clin Nutr Metab Care*. 2016. doi:10.1097/MCO.0000000000000254
22. Lattka E, Illig T, Heinrich J, Koletzko B. Do FADS genotypes enhance our knowledge about fatty acid related phenotypes? *Clin Nutr*. 2010. doi:10.1016/j.clnu.2009.11.005
23. Das UN. Nutritional factors in the prevention and management of coronary artery disease and heart failure. *Nutrition*. 2015. doi:10.1016/j.nut.2014.08.011
24. Martinelli N, Girelli D, Malerba G, et al. FADS genotypes and desaturase activity estimated

by the ratio of arachidonic acid to linoleic acid are associated with inflammation and coronary artery disease. *Am J Clin Nutr*. 2008;88(4):941-949. doi:10.1093/ajcn/88.4.941

25. Tosi F, Sartori F, Guarini P, Olivieri O, Martinelli N. Delta-5 and Delta-6 Desaturases: Crucial Enzymes in Polyunsaturated Fatty Acid-Related Pathways with Pleiotropic Influences in Health and Disease. In: ; 2014:61-81. doi:10.1007/978-3-319-07320-0_7
26. Sergeant S, Rahbar E, Chilton FH. Gamma-linolenic acid, Dihommo-gamma linolenic, Eicosanoids and Inflammatory Processes. *Eur J Pharmacol*. 2016. doi:10.1016/j.ejphar.2016.04.020
27. Fritsche KL. The Science of Fatty Acids and Inflammation. *Adv Nutr*. 2015. doi:10.3945/an.114.006940
28. Li SW, Wang J, Yang Y, et al. Polymorphisms in FADS1 and FADS2 alter plasma fatty acids and desaturase levels in type 2 diabetic patients with coronary artery disease. *J Transl Med*. 2016. doi:10.1186/s12967-016-0834-8
29. Cao Y, Lu L, Liang J, et al. Omega-3 Fatty Acids and Primary and Secondary Prevention of Cardiovascular Disease. *Cell Biochem Biophys*. 2015. doi:10.1007/s12013-014-0407-5
30. Altman DG, Royston P. The cost of dichotomising continuous variables. *BMJ*. 2006. doi:10.1136/bmj.332.7549.1080
31. Bush WS, Moore JH. Chapter 11: Genome-Wide Association Studies. *PLoS Comput Biol*. 2012. doi:10.1371/journal.pcbi.1002822
32. Kitsios GD, Dahabreh IJ, Trikalinos TA, Schmid CH, Huggins GS, Kent DM. Heterogeneity of the phenotypic definition of coronary artery disease and its impact on genetic association studies. *Circ Cardiovasc Genet*. 2011. doi:10.1161/CIRCGENETICS.110.957738
33. Polimanti R, Kranzler HR, Gelernter J. Phenome-wide association study for alcohol and nicotine risk alleles in 26394 women. *Neuropsychopharmacology*. 2016;41:2688-2696. doi:10.1038/npp.2016.72
34. Verma A, Lucas A, Verma SS, et al. PheWAS and Beyond: The Landscape of Associations with Medical Diagnoses and Clinical Measures across 38,662 Individuals from Geisinger. *Am J Hum Genet*. 2018;102(4):592-608. doi:10.1016/j.ajhg.2018.02.017
35. Karaca S, Civelek E, Karaca M, et al. Allergy-specific Phenome-Wide Association Study for Immunogenes in Turkish Children. *Sci Rep*. 2016;6:33152. doi:10.1038/srep33152
36. 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*. 2017;47(1):29. doi:10.1093/ije/dyx204
37. Lumley T. Analysis of Complex Survey Samples. *J Stat Softw*. 2004;9(8):1-19. doi:10.18637/jss.v009.i08