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