Lifestyle Risk Score for aggregating multiple lifestyle factors: Handling missingness of individual lifestyle components in meta-analysis of gene-by-lifestyle interactions

Hanfei Xu[1,*], Karen Schwander[2,*], Michael R Brown[3], Wenyi Wang[4], RJ Waken[5], Eric Boerwinkle[3,6], L Adrienne Cupples[1,7], Lisa de las Fuentes[8], Diana van Heemst[9], Oyomoare Osazuwa-Peters[10], Paul S de Vries[3], Ko Willems van Dijk[4,11,12], Yun Ju Sung[13], Xiaoyu Zhang[1], Alanna C Morrison[3], DC Rao[14], Raymond Noordam[9,#], Ching-Ti Liu[1,#]

[*] Shared-first author

[#] Shared-last author

1) Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA
2) Division of Statistical Genomics, Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA
3) Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, the University of Texas School of Public health, Houston, TX, USA
4) Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands
5) Field and Environmental Data Science, Benson Hill Inc, St. Louis, MO, USA
6) The Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas
7) NHLBI and Boston University Framingham Heart Study, Framingham, MA
8) Department of Medicine, Cardiovascular Division, Washington University School of Medicine, St. Louis, MO, USA
9) Section of Gerontology and Geriatrics, Department of Internal Medicine, Leiden University Medical Center, Leiden, the Netherlands
10) Department of Population Health Sciences, Duke University, Durham, NC, USA
11) Division of Endocrinology, Department of Internal Medicine, Leiden University Medical Center, Leiden, the Netherlands
12) Leiden Laboratory for Experimental Vascular Medicine, Leiden University Medical Center, Leiden, the Netherlands
13) Department of Psychiatry, Washington University School of Medicine, St. Louis, MO, USA
14) Division of Biostatistics, Washington University School of Medicine, St. Louis, MO, USA

Correspondence to ctliu@bu.edu (CTL) and hfxu@bu.edu (HX); Tel: 617-358-2482.

Abstract

Recent studies consider lifestyle risk score (LRS), an aggregation of multiple lifestyle exposures, in identifying association of gene-lifestyle interaction with disease traits. However, not all cohorts have data on all lifestyle factors, leading to increased heterogeneity in the environmental exposure in collaborative meta-analyses. We compared and evaluated four approaches (Naïve, Safe, Complete and Moderator Approaches) to handle the missingness in LRS-stratified meta-analyses under various scenarios. Compared to “benchmark” results with all lifestyle factors available for all cohorts, the Complete Approach, which included only cohorts with all lifestyle components, was underpowered, and the Naïve Approach, which utilized all available data and ignored the missingness, was slightly liberal. The Safe Approach, which used all data in LRS-exposed group and only included cohorts with all lifestyle factors available in the LRS-unexposed group, and the Moderator Approach, which handled missingness via moderator meta-regression, were both slightly conservative and yielded almost identical p-values. We also evaluated the performance of the Safe Approach under different scenarios. We observed that the larger the proportion of cohorts without missingness included, the more accurate the results compared to “benchmark” results. In conclusion, we generally recommend the Safe Approach to handle heterogeneity in the LRS based genome-wide interaction meta-analyses.

1. Introduction

Thanks to strong collaborations, many large-scale genome-wide association studies (GWAS) have successfully identified many genetic determinants described to explain part of the pathophysiological mechanism underlying a wide range of traits. Despite these efforts and increased sample sizes, the explained variability of many traits is relatively small and only a small proportion of the familial heritability can be explained by the candidate variants found (Evangelou et al., 2018; López-Cortegano & Caballero, 2019; Manolio et al., 2009).

In addition to genetics, environmental factors and gene-environment interactions may contribute to this unexplained trait heritability (Manolio et al., 2009; Rao et al., 2017). Recently, genome-wide gene-environment interaction studies have been conducted to further explore the potential mechanisms underlying an array of diseases or disease traits of interest (de las Fuentes et al., 2020; Graff et al., 2017; Liu et al., 2012; Noordam et al., 2019; Wu et al., 2020). Thus far, these collaborative efforts have largely focused on a single environmental or lifestyle factor, such as smoking (Bentley et al., 2019; Justice et al., 2017; Yun J. Sung et al., 2018; Yun Ju Sung et al., 2019; Wu et al., 2020), physical activity (Graff et al., 2017; Kilpeläinen et al., 2019), alcohol intake (De Vries et al., 2019), educational attainment (de las Fuentes et al., 2020) and others (Jiang et al., 2018; Noordam et al., 2019). By accounting for the environmental risk factor, these efforts identified several novel loci beyond those identified by the traditional main effects-only GWAS. However, multiple environmental factors may simultaneously modify the genetics effects of loci (Osazuwa-Peters et al., 2020). Additionally, single lifestyle variables may not capture the spectrum of relevant environmental variation, resulting in biased effect estimation and false-negative results due to reduced statistical power.

Lifestyle factors, such as smoking, physical inactivity and alcohol consumption, all contribute independently to the risk of developing cardiovascular diseases, and composite lifestyle risk scores (LRS) have been used previously to assess the combined effect of multiple lifestyle factors on cardiovascular disease development (Abdullah Said, Verweij, & Van Der Harst, 2018; Lévesque, Poirier, Després, & Alméras, 2017; Sotos-Prieto, Baylin, Campos, Qi, & Mattei, 2016). However, when applying LRS methodology to large collaborative consortium settings, challenges arise as not all lifestyle components in the LRS are available in all participating cohorts and/or may not be measured using the same instrument. If ignored, significant measurement error and potential heterogeneity may be introduced with reduced statistical power and potential bias. In the present study, we explore different approaches for incorporating missingness of individual lifestyle components with meta-analysis of genome-wide gene-environment interaction on systolic blood pressure in four European-ancestry (EA) cohorts.

2. Methods

2.1 Participating cohorts and subject inclusion

In this study, we included data from four cohorts, which were the Atherosclerosis Risk in Communities Study (ARIC), the Framingham Heart Study (FHS), the Hypertension Genetic Epidemiology Network (HyperGEN), and the Netherlands Epidemiology of Obesity Study (NEO). For cohorts with data collected from multiple center visits, we chose a single visit that could maximize sample size with non-missing data. We included a total of 24,048 EA individuals who were aged 18-80 and had non-missing genotype, phenotype and relevant covariates information, including age, sex, systolic blood pressure (SBP), anti-hypertensive medications, body mass index (BMI) and the four lifestyle factors (smoking status, alcohol consumption, education level, and physical activity).

2.2 Phenotype and covariates

Resting SBP (mmHg) was calculated by taking the average of all available BP readings at the same clinical visit, and further adjusted by adding 15 mmHg for subjects with anti-hypertensive medication use (Tobin, Sheehan, Scurrah, & Burton, 2005). SBP values that were more than six standard deviations away from the mean were winsorized to exactly at six standard deviations from the mean, in order to reduce the potential influence of outliers.

Other covariates included age, sex, field center (if appropriate), and principal components to account for population stratification. Analyses were performed with and without further adjusting for BMI.

2.3 Genotyping and QC

Genotyping was performed separately within each cohort using Affymetrix (Santa Clara, CA, USA) or Illumina (San Diego, CA, USA) genotyping arrays (Supplementary Table S1). Each cohort performed imputations with IMPUTE2 (Howie, Donnelly, & Marchini, 2009) or MaCH (Li, Willer, Ding, Scheet, & Abecasis, 2010), using the cosmopolitan reference panel from the 1000 Genomes

Project Phase 1 Integrated Release Version 3 Haplotypes (2010-11 data freeze, 2012-03-14 haplotypes) (Altshuler et al., 2012). SNPs were excluded if they were non-autosomal, had minor allele frequency (MAF) <1% or low imputation quality (r^2<0.1). We conduct further quality control filters centrally during the meta-analysis (Section 2.5.2).

2.4 Lifestyle Risk Score

In this study, we considered four lifestyle factors: smoking status (never/former/current smoker), current alcohol intake (drinks per week), educational attainment beyond high school (none/some college/college degree) and physical activity (inactive/active). We classified participants as “college degree” if they completed at least a 4-year college degree, as “some college” if they received any education beyond high school including vocational school but did not complete a college degree and as “none” if they received no education beyond high school (de las Fuentes et al., 2020). Physical activity is expressed in metabolic equivalents (MET; 1 MET = 1 kcal/kg/hour). Inactive individuals were defined as those with <225 MET- minutes per week of moderate-to-vigorous leisure-time or commuting physical activity, or in the lower quartile (25%) of the physical activity distribution within cohort. The detailed definitions of active and inactive physical activity followed a previous study on gene-physical activity interaction (Kilpeläinen et al., 2019).

Construction of the lifestyle risk score can be separated into two steps. First, each lifestyle factor, treated as an individual lifestyle component, was categorized into no risk (with value of 0), low risk (with value of 1) and high risk (with value of 2) based on its effect on BP or cardiovascular health, except physical activity which only had no risk and low risk (Osazuwa-Peters et al., 2020).

The higher risk value the category was assigned, the more relevant to unfavorable cardiovascular health outcomes. Note that we categorized modest alcohol intake as no risk and abstinence as low risk because there was evidence that moderate alcohol consumption had consistently been associated with a decreased risk of type 2 diabetes (Joosten et al., 2010) and coronary artery disease (Klatsky, 1999) compared with abstention or excessive consumption (Feitosa et al., 2018). Table 1 provides the details of lifestyle risk score component definition.

Second, the "Complete" Quantitative Lifestyle Risk Score (QLRS-C) was calculated by summing up all four components, ranging from 0-7. We also calculated the "Partially Missing" Quantitative Lifestyle Risk Score (QLRS-M) using 2-3 components pre-selected for each cohort by design, as described in Table 2. For example, for ARIC, we included three lifestyle components (smoking, education and physical activity) when constructing QLRS-M. QLRS-M ranges from 0 to 4 or 5, depending on the inclusion of lifestyle components for each cohort.

After constructing the Quantitative Lifestyle Risk Scores, we further created Dichotomous Lifestyle Risk Scores for the "Complete" (DLRS-C) and the "Partial" (DLRS-M) summary scores. We gave a value of 0 (unexposed group) if the corresponding Quantitative Lifestyle Risk Score <2 and a value of 1 (exposed group) if Quantitative Lifestyle Risk Score ≥2 (i.e. at least one risk component classified as high risk or at least two components classified as low risk). These dichotomized LRS measures are used to define exposed and unexposed strata in our analyses.

It is worth noting that cohorts with partially missing lifestyle components have equal or lower LRS than its actual score had we observed all lifestyle components. This leads to potential misclassification when dichotomizing the LRS into exposed and unexposed groups. However, no participant would be misclassified as exposed and they can only be misclassified as unexposed, leading to heterogeneity in the unexposed group only.

2.5 Statistical Analysis

2.5.1 Overview

We conduct a two-stage analysis procedure. In Stage 1, each cohort performed LRS-stratified genome-wide association analysis on SBP using the main effect model ($E(Y) = \beta_0 + \beta_G\, SNP + \beta_C$ *Covariates*, where Y is the SBP level, SNP is the imputed additive dosage value of the genetic variant), in DLRS-C exposed and DLRS-C unexposed strata. The association analyses were also repeated in DLRS-M exposed and DLRS-M unexposed strata. In Stage 2, we performed meta-analysis within each stratum, and then evaluated the joint effects of main and interaction effects by calculating the p-values for the 2 degree of freedom joint test. Under Stage 2, we considered four different meta-analysis approaches of handling missingness of lifestyle components (Na ïve, Safe, Complete and Moderator Approaches). We evaluated the performance of the four approaches under four scenarios where some cohorts were designed to provide association results using “Complete” LRS but the others were designed to only provide “Partially Missing” results.

2.5.2 Stage 1: Cohort-specific stratified analysis and QC of association results

For Stage 1, each cohort performed eight genome-wide association analyses on SBP using the main effect model: two strata (exposed/unexposed) × two LRS (DLRS-C/DLRS-M) × two BMI adjustment (with/without). Association analyses were implemented either using ProbABEL (Aulchenko, Struchalin, & van Duijn, 2010) for studies with unrelated samples, or using MMAP (https://mmap.github.io/) for studies with family relatedness. Each cohort provided the robust estimates of the stratum-specific genetic main effect and corresponding robust standard error (SE) for all eight analyses. Cohort-specific details are presented in Supplementary Table S1.

We performed extensive quality control (QC) using the R package EasyQC (Winkler et al., 2014) on each of the eight cohort-specific association results, which contained approximately 8-9 million variants. First, we removed variants with invalid alleles and indels, harmonized alleles and variant names across cohorts, and compared allele frequencies with the ancestry-specific 1000 Genomes reference panel. Next, we compared summary statistics (e.g., mean, standard deviation, minimum, maximum) of estimated effect sizes, standard errors, and p-values across cohorts to identify potential outliers, and reviewed SE-N (i.e., inverse of the median standard error versus the square root of the sample size) plots to look for possible problems with phenotype or covariates (Winkler et al., 2014). Finally, SNPs were excluded if imputation quality score was <0.5, or if the product of the imputation quality and minor allele count was <20 (de las Fuentes et al., 2020; Kilpeläinen et al., 2019). No genomic control was applied after filtering as there were little to no problems with inflation (genomic control inflation λ ranged from 1.020 to 1.071).

Since QC and filtering were performed separately within each stratum, the set of variants remaining in each stratum differed slightly. Thus, we further harmonized the set of variants between the exposed and unexposed strata within each LRS construction – BMI adjustment combination, to ensure that the set of variants was identical between strata. After QC, the number of variants in each association result was between 5.3M-8.2M.

2.5.3 Stage 2: Meta-analysis

After obtaining cohort-specific GWAS results using “Complete” and “Partially Missing” LRS, we first performed meta-analyses within each stratum (exposed/unexposed) using the results obtained from analyses using “Complete” LRS, and considered this set of meta-analyzed results as the “benchmark” results as there is no missing lifestyle component in each cohort’s LRS construction.

Then, to mimic the real life situation where some of the cohorts would provide GWAS association results obtained from analyses using “Complete” LRS but the others could only provide “Partially Missing” results, we further performed the meta-analyses using a mixture of results obtained from cohort-specific analyses conducted with “Complete” LRS and “Partially Missing” LRS. We considered four scenarios using different cohort mixture patterns by changing each cohort’s contribution of lifestyle components, in order to better utilize the data. The setting of each scenario is presented in Table 3. For example, Scenario 1 is to use “complete” results from ARIC, and “partially missing” results from HyperGEN, FHS and NEO.

As mentioned in the LRS section, the missingness in lifestyle components will cause misclassification when dichotomizing LRS into exposed and unexposed groups, hence leading to heterogeneity in the unexposed group only. To account for this heterogeneity, we considered four different meta-analysis approaches of utilizing “Complete” and “Partially Missing” results under various scenarios discussed above.

1) Naïve Approach. This approach simply takes all association results contributed by each participating cohort without worrying whether their LRS includes all lifestyle components, for both exposed and unexposed groups.

2) Safe Approach. Since heterogeneity only occurs in the unexposed group, it is "safe" to only take association results from cohorts with complete LRS for the unexposed group analysis, while including results from all cohorts no matter whether the missing data exist in LRS for the exposed group analysis.
3) Complete Approach. This approach only uses association results from cohorts with complete LRS data in meta-analysis, for both exposed and unexposed groups.
4) Moderator Approach. This approach uses all the contributed data from cohorts without regard to their missingness in lifestyle components. It utilizes the framework of meta-regression, while including moderator terms indicating the missing LRS components across cohorts in the design matrix of the meta-regression to account for missingness during meta-analysis. Technical details of this approach are available in the Supplementary Method.

Table 3 also shows the inclusion of association results in the meta-analysis using each of the approaches described above under Scenarios 1-4. Here we take Scenario 1 as an example: For the Naïve Approach, we analyze exposed and unexposed groups separately using "complete" results from ARIC, and "partially missing" results from HyperGEN, FHS and NEO without differentiating "complete" or "partially missing". For the Safe Approach, we include ARIC results alone and ignore other cohorts' contributions with "partially missing" results for the unexposed group; for the exposed group, we analyze all four cohorts using "complete" results from ARIC, and "partially missing" results from HyperGEN, FHS and NEO. For the Complete Approach, we analyze exposed and unexposed groups separately, but only use "complete" results from ARIC with no other cohorts included. For the Moderator Approach, we take "complete" results from

ARIC, and “partially missing” results from HyperGEN, FHS and NEO for both exposed and unexposed groups as input of the meta-regression.

For the “benchmark” meta-analysis and the first three approaches (Na ïve, Safe and Complete), we used METAL software (Willer, Li, & Abecasis, 2010) to perform meta-analyses within each stratum and used EasyStrata (Winkler et al., 2015) to calculate the 2 degree of freedom joint p-values. For the Moderator Approach, we used the Moderator Web App and R code developed by Dr. RJ Waken (https://rjwaken.shinyapps.io/missing_lrs_meta/).

In the following sections, we focus on the comparison of results obtained from the analyses without adjusting for BMI, since we observed the same pattern for BMI-adjusted analyses and the primary objective of our study is to evaluate the meta-analysis approaches of handling missingness instead of identifying novel loci under confounder adjustment.

3. Results

3.1 Sample Characteristics

Sample characteristics are presented in Supplementary Tables S2, S3, and S4. ARIC had the largest sample size (N=9,426) and HyperGen cohort had the fewest number of participants (N=1,249). All cohorts had similar distributions of sex, age and BMI, except that FHS and HyperGEN had a wider age range than ARIC and NEO (Supplementary Table S2). In Supplementary Tables S3 and S4, the exposed group had slightly higher SBP level than the unexposed group for all four cohorts in terms of DLRS-C. However, the difference in SBP levels between exposed and unexposed groups was smaller when we defining exposure groups using DLRS-M. The proportion of subjects in the exposed group was smaller when using DLRS-M compared to DLRS-C, indicating potential misclassification.

3.2 Results Comparison between Approaches

Figure 1 presents the results of comparison of the four meta-analysis approaches to the "benchmark" results derived from analyses of cohorts where all lifestyle components were present. Among variants that reach genome-wide significance level ($p\text{-value}<5\times10^{-8}$), we observed that the Complete Approach yielded much larger p-values than the "benchmark" results, thus could be considered underpowered. The Na ïve Approach was able to detect the same set of genome-wide significant variants as the "benchmark" results, but with slightly smaller p-values. The Safe and Moderator Approaches led to slightly larger p-values than "benchmark" results. The Q-Q plot (Figure 2) also shows that the Complete Approach obtained the most deflated p-values among the four approaches ($\lambda_{\text{Complete vs benchmark}} = 0.972$). The Safe Approach and Moderator Approach yielded similar slightly conservative results ($\lambda_{\text{Safe vs benchmark}} = \lambda_{\text{Moderator vs benchmark}} =0.985$), while the results of the Na ïve Approach were slightly liberal ($\lambda_{\text{Naive vs benchmark}} = 1.004$).

Figure 3 shows the pair-wise comparison among four meta-analysis approaches. Similar to what we observed in Figure 1, the Complete Approach was underpowered compared to other three approaches. The Safe and Moderator Approaches yielded similar but slightly larger p-values than Na ïve Approach, and the degree of similarity increased with significance. Notably, the results of Safe Approach and Moderator Approach were almost identical, but the number of variants included in the analyses for the Moderator Approach (Number of variants = 5,258,666) was much smaller than the Safe Approach (Number of variants =8,181,669).

Scenarios 2-4 presented similar patterns to Scenario 1 in terms of comparison with "benchmark" results and within-scenario comparison between different approaches; so we focus on illustrating Scenario 1 in the results section. The detailed comparison results of Scenarios 2-4 are available in the Supplementary Figures S1-S9.

3.3 Result Comparison between Scenarios for Safe Approach

Here we further evaluated the performance of the same meta-analysis approach under different scenarios. Since we generally were concerned with false-positive results, we focused our attention only to the non-inflated Safe Approach. Figure 4 presents the scatterplot of association results between "benchmark" and the Safe Approach for each of the four scenarios, for variants with p-value$<1\times10^{-6}$ in at least one of comparing results. We observed that for SNPs reaching genome-wide significance (p-value$<5\times10^{-8}$) in "benchmark" results, the points of Scenarios 3 and 4 almost lay along the diagonal line, while points of Scenarios 1 and 2 were a bit away from the diagonal. This indicated that the Safe Approach under Scenarios 3 and 4 more accurately identified positive signals than under Scenarios 1 and 2.

The Q-Q plot (Supplementary Figure S10) shows that when p-values were large ($>10^{-5}$), Scenario 4 with less missingness provided more similar p-value distributions with "benchmark" results ($\lambda_{\text{scenario 4 vs benchmark}}$=0.991) compared to Scenario 1 ($\lambda_{\text{scenario 1 vs benchmark}}$=0.983) and Scenario 3 ($\lambda_{\text{scenario 3 vs benchmark}}$=0.984). Although Scenario 2 seemed to perform very well on large p-values

($\lambda_{\text{scenario 2 vs benchmark}}$=0.994), it provided substantially deflated results toward the tail when reaching genome-wide significance. In the meantime, Scenarios 3 and 4 had similar p-value distributions and both of their p-values were very close to the "benchmark" distribution when p-values were small. The p-values of Scenario 1 were closer to the diagonal line than those of Scenario 2 when p-values were small, and this may due to the sample size of the cohort with "complete" results in Scenario 1 (ARIC, N=9,426), which was greater than that of Scenario 2 (FHS, N=7,638).

In general, we consider Scenario 4 performed better than Scenario 3, in turn better than Scenarios 1 and 2. This meets our expectation as Scenario 4 had the smallest proportion of cohorts using "partially missing" results; thus it was expected to bring the most comprehensive information into meta-analysis.

4. Discussion

In this study, we evaluated four different strategies handling the missingness of individual lifestyle components in the meta-analysis of gene-lifestyle interaction using LRS-stratified summary statistics from participating cohorts. We aimed to find the best way to leverage the available data while appropriately handling the heterogeneity due to missing data in the LRS, and further improve the power of identifying novel loci for the trait of interest. Only utilizing data contributed by the cohorts without missingness in any lifestyle components (the Complete Approach) is very underpowered, while freely meta-analyzing all the association results contributed by the cohorts even with missing components in the LRS (the Naïve Approach) is slightly liberal. The Safe Approach and Moderator Approach are both slightly conservative and their p-values are almost identical to each other. We also observed that, as expected, the more cohorts with information for all lifestyle components we used in meta-analysis, the more accurate the results. This result confirms our primary hypothesis.

A risk score is a commonly used approach to evaluate combined effects of risk factors and it may play an important role in personalized medicine. In the past, the scientific community has proposed several well-known risk scores. For example, the Framingham Risk Score (Wilson et al., 1998) is a sex-specific score used to estimate the 10-year cardiovascular risk, and the diabetes risk score (Lindström & Tuomilehto, 2003) is a screening tool for identifying subjects at high-risk for type 2 diabetes. Lifestyle Risk Scores have also become popular as people are increasingly interested in their clinical implications drawn by the joint effects of individual lifestyle factors to a specific trait, disease, or time-to event outcome. In the meantime, the genetic risk score (GRS) has become a widely used tool to improve identification of persons who are at risk for common complex diseases after numerous stories about exceptional success in genome-wide association studies (GWAS).

There have been some prior studies combining genetic risk scores and lifestyle risk scores to explore their joint behavior on risk of CVD (Abdullah Said et al., 2018) and Colorectal Cancer (Cho et al., 2019). Specifically, these studies divided study samples into subgroups based on the combination of genetic risk score level and lifestyle risk score level, and found that within and across genetic risk groups, adherence to poor behavioral lifestyle was associated with increased risk of diseases, and there was no interaction effect between genetic risk and lifestyle risk. This might seem discouraging whether adding genetic information could add much to the risk prediction studies using lifestyle risk scores. However, it is important to note that the genetic risk score was calculated based on variants reported from previous standard genome-wide significant analyses without taking its potential modification effect into consideration; variants whose effects may differ by level of lifestyle risk score might therefore be missed by standard GWAS screening. Moreover, a LRS may have a different modification effect on each variant, so instead of looking at aggregated genetic risk score only, interaction with one variant at a time should also be evaluated. Our study looked into the combination of genetic and lifestyle information in the way of performing meta-analysis of gene-by-lifestyle interaction in order to find novel loci for complex disease traits, and those potential novel loci may provide additional information for computing a genetic risk score, which could increase the power of previous studies.

Handling missing data in the aggregation of risk factors is challenging, yet important and worth the effort to explore in further detail. Based on the properties of genetic architecture, GRS can be computed using imputed or proxy SNPs, when the originally reported variants are not available, based on the largely available reference panel, such as 1000 Genome Project (Auton et al., 2015). Thus, it is more flexible than LRS in terms of dealing with missingness. There were several methods proposed to impute phenotypes using the correlation structure between phenotypes, family structure or information from other cohorts (Chen, Peloso, & Dupuis, 2018; Dahl et al., 2016; Hormozdiari et al., 2016), but these methods rarely dealt with the case that one phenotype is completely unavailable for all the individuals in one particular cohort contributing to a large meta-analysis, which is what we encountered in our study. When considering using summary statistics in meta-analysis, a previous study (Loef & Walach, 2012) tried to deal with the issue of missingness by restricting the study sample to cohorts with at least three out of five lifestyle behaviors available, reducing sample size and thus power to a great extent, with the issue of heterogeneity unresolved. Our study proposes making the best use of the available data gathered from cohorts to obtain accurate combined effects of risk factors, thereby providing a novel perspective for risk score based meta-analysis in future research.

Our study examined the Moderator Approach, which is a novel approach of accounting for missingness via meta-regression in the gene-by-environment interaction field. Instead of performing stratum-specific meta-analyses and then evaluating the interaction, this approach can achieve the final goal in one step via meta-regression, with meta-analysis results of both exposure groups as input. However, due to the meta-regression setting, the Moderator Approach requires that the number of cohorts have GWAS results available for a SNP (which equals to 4 in our study) is greater than the number of predictors divided by two (which equals to [one main effect + one interaction effect + four missingness effects]/2 = 3 in our study). Therefore, it restricted the analyses to the SNPs existing in the GWAS results of all four cohorts, thereby eliminating a large number of SNPs from the analyses and possibly missed positive signals. On the other hand, the design matrix of the meta-regression model in the Moderator Approach should be treated with caution because in some cases of missingness pattern, the design matrix would suffer from multicollinearity and we could not successfully obtain the least square estimates. Since the Safe Approach can provide almost identical results as the Moderator Approach but does not have a restriction on the missingness pattern and the number of cohorts and predictors, we would recommend using the Safe Approach to handle missingness during meta-analysis. Potential future directions of our study would be to further investigate the Moderator approach and to evaluate the performance of Safe Approach and Moderator Approach under larger scale of meta-analyses.

Strengths and Limitations

Our study has several important strengths. To our knowledge, this is the very first study to explore how to deal with missingness in individual lifestyle components in order to improve the power for identifying novel genetic loci for complex disease traits. Our study performed thorough comparisons between four meta-analysis approaches via various cohort mixture scenarios, thus providing comprehensive information for investigators to refer to.

Although this study has several strengths as an innovative work for dealing with missingness in gene-by-lifestyle interaction, it has some limitations. When calculating the “Partially Missing” LRS, we assigned the missingness pattern to each of the cohorts. Also, when performing the meta-analyses using a mixture of results obtained from cohort-specific analysis conducted with “Complete” LRS or “Partially Missing” LRS, although we considered various cohort mixture scenarios, we still were not able to catch every possible pattern. This kind of design may lose some flexibility and consequently fail to capture all the information during the comparison. Moreover, our study mainly evaluated the performance of different approaches in terms of joint effects instead of focusing on the interaction effect. We did not manage to capture a clear pattern when comparing the interaction effect between different meta-analysis approaches, due to the small sample size of our study. It is worth pursuing the comparison of the interaction effect itself among different approaches with a larger sample size, by incorporating more cohorts in our next step.

In summary, we evaluated four approaches of incorporating the missingness of lifestyle components in the meta-analysis of gene-by-lifestyle interaction. Based on our results, we generally recommend using the Safe Approach since it is straightforward to implement and yields non-inflated results. Handling the missingness of individual lifestyle components properly can efficiently increase statistical power of gene-by-lifestyle interaction analysis for identifying novel loci of complex traits.

Acknowledgements:

This project was largely supported by a grant from the U.S. National Heart, Lung, and Blood Institute (NHLBI), the National Institutes of Health, R01HL118305. The full list of acknowledgments appears in the Supplementary Note.

Conflict of Interest Statement:

The authors declare that there is no conflict of interests.

References:

Abdullah Said, M., Verweij, N., & Van Der Harst, P. (2018). Associations of combined genetic and lifestyle risks with incident cardiovascular disease and diabetes in the UK biobank study. *JAMA Cardiology*, *3*(8), 693–702. https://doi.org/10.1001/jamacardio.2018.1717

Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., … Lacroute, P. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56–65. https://doi.org/10.1038/nature11632

Aulchenko, Y. S., Struchalin, M. V., & van Duijn, C. M. (2010). ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics*, *11*(1), 134. https://doi.org/10.1186/1471-2105-11-134

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., … Schloss, J. A. (2015, September 30). A global reference for human genetic variation. *Nature*. Nature Publishing Group. https://doi.org/10.1038/nature15393

Bentley, A. R., Sung, Y. J., Brown, M. R., Winkler, T. W., Kraja, A. T., Ntalla, I., … Cupples, L. A. (2019). Multi-ancestry genome-wide gene–smoking interaction study of 387,272 individuals identifies new loci associated with serum lipids. *Nature Genetics*, *51*(4), 636–648. https://doi.org/10.1038/s41588-019-0378-y

Chen, Y., Peloso, G. M., & Dupuis, J. (2018). Evaluation of a phenotype imputation approach using GAW20 simulated data 01 Mathematical Sciences 0104 Statistics. In *BMC Proceedings* (Vol. 12, p. 56). BioMed Central Ltd. https://doi.org/10.1186/s12919-018-0134-9

Cho, Y. A., Lee, J., Oh, J. H., Chang, H. J., Sohn, D. K., Shin, A., & Kim, J. (2019). Genetic

Risk Score, Combined Lifestyle Factors and Risk of Colorectal Cancer. *Cancer Research and Treatment*, *51*(3), 1033–1040. https://doi.org/10.4143/crt.2018.447

Dahl, A., Iotchkova, V., Baud, A., Johansson, S., Gyllensten, U., Soranzo, N., … Marchini, J. (2016). A multiple-phenotype imputation method for genetic studies. *Nature Genetics*, *48*(4), 466–472. https://doi.org/10.1038/ng.3513

DAWBER, T. R., MEADORS, G. F., & MOORE, F. E. (1951). Epidemiological approaches to heart disease: the Framingham Study. *American Journal of Public Health*, *41*(3), 279–281. https://doi.org/10.2105/AJPH.41.3.279

de las Fuentes, L., Sung, Y. J., Noordam, R., Winkler, T., Feitosa, M. F., Schwander, K., … Fornage, M. (2020). Gene-educational attainment interactions in a multi-ancestry genome-wide meta-analysis identify novel blood pressure loci. *Molecular Psychiatry*. https://doi.org/10.1038/s41380-020-0719-3

De Vries, P. S., Brown, M. R., Bentley, A. R., Sung, Y. J., Winkler, T. W., Ntalla, I., … Morrison, A. C. (2019). Multiancestry Genome-Wide Association Study of Lipid Levels Incorporating Gene-Alcohol Interactions. *American Journal of Epidemiology*, *188*(6), 1033–1054. https://doi.org/10.1093/aje/kwz005

Evangelou, E., Warren, H. R., Mosen-Ansorena, D., Mifsud, B., Pazoki, R., Gao, H., … Caulfield, M. J. (2018). Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. *Nature Genetics*, *50*(10), 1412–1425. https://doi.org/10.1038/s41588-018-0205-x

Feitosa, M. F., Kraja, A. T., Chasman, D. I., Sung, Y. J., Winkler, T. W., Ntalla, I., … Levy, D. (2018). Novel genetic associations for blood pressure identified via gene-alcohol interaction in up to 570K individuals across multiple ancestries. *PLoS ONE*, *13*(6). https://doi.org/10.1371/journal.pone.0198166

Graff, M., Scott, R. A., Justice, A. E., Young, K. L., Feitosa, M. F., Barata, L., … Kilpeläinen, T. O. (2017). Genome-wide physical activity interactions in adiposity — A meta-analysis of 200,452 adults. *PLoS Genetics*, *13*(4). https://doi.org/10.1371/journal.pgen.1006528

Hormozdiari, F., Kang, E. Y., Bilow, M., Ben-David, E., Vulpe, C., McLachlan, S., … Eskin, E. (2016). Imputing Phenotypes for Genome-wide Association Studies. *American Journal of Human Genetics*, *99*(1), 89–103. https://doi.org/10.1016/j.ajhg.2016.04.013

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genetics*, *5*(6), e1000529. https://doi.org/10.1371/journal.pgen.1000529

Jiang, X., O’Reilly, P. F., Aschard, H., Hsu, Y. H., Richards, J. B., Dupuis, J., … Kiel, D. P. (2018). Genome-wide association study in 79,366 European-ancestry individuals informs the genetic architecture of 25-hydroxyvitamin D levels. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-017-02662-2

Joosten, M. M., Grobbee, D. E., van der A, D. L., Verschuren, W. M., Hendriks, H. F., & Beulens, J. W. (2010). Combined effect of alcohol consumption and lifestyle behaviors on risk of type 2 diabetes. *The American Journal of Clinical Nutrition*, *91*(6), 1777–1783. https://doi.org/10.3945/ajcn.2010.29170

Justice, A. E., Winkler, T. W., Feitosa, M. F., Graff, M., Fisher, V. A., Young, K., … Cupples, L. A. (2017). Genome-wide meta-analysis of 241,258 adults accounting for smoking behaviour identifies novel loci for obesity traits. *Nature Communications*, *8*.

https://doi.org/10.1038/ncomms14977

Kannel, W. B., Feinleib, M., McNamara, P. M., Garrison, R. J., & Castelli, W. P. (1979). An investigation of coronary heart disease in families. *American Journal of Epidemiology*, *110*(3), 281–290. https://doi.org/10.1093/oxfordjournals.aje.a112813

Kilpeläinen, T. O., Bentley, A. R., Noordam, R., Sung, Y. J., Schwander, K., Winkler, T. W., … Loos, R. J. F. (2019). Multi-ancestry study of blood lipid levels identifies four loci interacting with physical activity. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-018-08008-w

Klatsky, A. L. (1999). Moderate drinking and reduced risk of heart disease. *Alcohol Research and Health*, *23*(1), 15–22.

Lévesque, V., Poirier, P., Després, J. P., & Alméras, N. (2017). Relation Between a Simple Lifestyle Risk Score and Established Biological Risk Factors for Cardiovascular Disease. *American Journal of Cardiology*, *120*(11), 1939–1946. https://doi.org/10.1016/j.amjcard.2017.08.008

Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010). MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, *34*(8), 816–834. https://doi.org/10.1002/gepi.20533

Lindström, J., & Tuomilehto, J. (2003). The diabetes risk score: A practical tool to predict type 2 diabetes risk. *Diabetes Care*, *26*(3), 725–731. https://doi.org/10.2337/diacare.26.3.725

Liu, C. T., Estrada, K., Yerges-Armstrong, L. M., Amin, N., Evangelou, E., Li, G., … Hsu, Y. H. (2012). Assessment of gene-by-sex interaction effect on bone mineral density. *Journal of Bone and Mineral Research*, *27*(10), 2051–2064. https://doi.org/10.1002/jbmr.1679

Loef, M., & Walach, H. (2012, September 1). The combined effects of healthy lifestyle behaviors on all cause mortality: A systematic review and meta-analysis. *Preventive Medicine*. Academic Press. https://doi.org/10.1016/j.ypmed.2012.06.017

López-Cortegano, E., & Caballero, A. (2019). Inferring the nature of missing heritability in human traits using data from the GWAS catalog. *Genetics*, *212*(3), 891–904. https://doi.org/10.1534/genetics.119.302077

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., … Visscher, P. M. (2009, October 8). Finding the missing heritability of complex diseases. *Nature*. Nature Publishing Group. https://doi.org/10.1038/nature08494

Noordam, R., Bos, M. M., Wang, H., Winkler, T. W., Bentley, A. R., Kilpeläinen, T. O., … Redline, S. (2019). Multi-ancestry sleep-by-SNP interaction analysis in 126,926 individuals reveals lipid loci stratified by sleep duration. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-12958-0

Osazuwa-Peters, O. L., Waken, R. J., Schwander, K. L., Sung, Y. J., de Vries, P. S., Hartz, S. M., … Rao, D. C. (2020). Identifying blood pressure loci whose effects are modulated by multiple lifestyle exposures. *Genetic Epidemiology*. https://doi.org/10.1002/gepi.22292

Rao, D. C., Sung, Y. J., Winkler, T. W., Schwander, K., Borecki, I., Adrienne Cupples, L., … Psaty, B. M. (2017). Multiancestry Study of Gene-Lifestyle Interactions for Cardiovascular Traits in 610 475 Individuals from 124 Cohorts: Design and Rationale. *Circulation: Cardiovascular Genetics*, *10*(3). https://doi.org/10.1161/CIRCGENETICS.116.001649

Sotos-Prieto, M., Baylin, A., Campos, H., Qi, L., & Mattei, J. (2016). Lifestyle Cardiovascular Risk Score, Genetic Risk Score, and Myocardial Infarction in Hispanic/Latino Adults

Living in Costa Rica. *Journal of the American Heart Association*, *5*(12). https://doi.org/10.1161/JAHA.116.004067

Sung, Yun J., Winkler, T. W., de las Fuentes, L., Bentley, A. R., Brown, M. R., Kraja, A. T., … Chasman, D. I. (2018). A Large-Scale Multi-ancestry Genome-wide Study Accounting for Smoking Behavior Identifies Multiple Significant Loci for Blood Pressure. *American Journal of Human Genetics*, *102*(3), 375–400. https://doi.org/10.1016/j.ajhg.2018.01.015

Sung, Yun Ju, de las Fuentes, L., Winkler, T. W., Chasman, D. I., Bentley, A. R., Kraja, A. T., … Morrison, A. C. (2019). A multi-ancestry genome-wide study incorporating gene–smoking interactions identifies multiple new loci for pulse pressure and mean arterial pressure. *Human Molecular Genetics*, *28*(15), 2615–2633. https://doi.org/10.1093/hmg/ddz070

THE ARIC INVESTIGATORS. (1989). THE ATHEROSCLEROSIS RISK IN COMMUNITIES (ARIC) STUDY: DESIGN AND OBJECTIVES. *American Journal of Epidemiology*, *129*(4), 687–702. https://doi.org/10.1093/oxfordjournals.aje.a115184

Tobin, M. D., Sheehan, N. A., Scurrah, K. J., & Burton, P. R. (2005). Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure. *Statistics in Medicine*, *24*(19), 2911–2935. https://doi.org/10.1002/sim.2165

Willer, C. J., Li, Y., & Abecasis, G. R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *BIOINFORMATICS APPLICATIONS NOTE*, *26*(17), 2190–2191. https://doi.org/10.1093/bioinformatics/btq340

Wilson, P. W. F., D’Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., & Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation*, *97*(18), 1837–1847. https://doi.org/10.1161/01.CIR.97.18.1837

Winkler, T. W., Day, F. R., Croteau-Chonka, D. C., Wood, A. R., Locke, A. E., Mägi, R., … Loos, R. J. F. (2014). Quality control and conduct of genome-wide association meta-analyses. *Nature Protocols*, *9*(5), 1192–1212. https://doi.org/10.1038/nprot.2014.071

Winkler, T. W., Kutalik, Z., Gorski, M., Lottaz, C., Kronenberg, F., & Heid, I. M. (2015). EasyStrata: evaluation and visualization of stratified genome-wide association meta-analysis data. *Bioinformatics (Oxford, England)*, *31*(2), 259–261. https://doi.org/10.1093/bioinformatics/btu621

Wu, P., Rybin, D., Bielak, L. F., Feitosa, M. F., Franceschini, N., Li, Y., … Vassy, J. L. (2020). Smoking-by-genotype interaction in type 2 diabetes risk and fasting glucose. *PLOS ONE*, *15*(5), e0230815. https://doi.org/10.1371/journal.pone.0230815

Table 1. Definition of Lifestyle Risk Score Component, with no risk as the value of 0, low risk as the value of 1 and high risk as the value of 2.

Component variables	No risk (0)	Low risk (1)	High risk (2)
Smoking	Never	Former	Current
Current Alcohol Intake (drinks/week)	Modest (1 – 7)	Abstinence* (0)	Heavy (> 7)
Education (after High School)	College Degree	Some College	None
Physical Activity	Active	Inactive	--

** Includes Former Drinkers*

Table 2. Components included in the calculation of "Partially Missing" Lifestyle Risk Score (QLRS-M) for each cohort (by design)

Cohort	Smoking	Alcohol	Education	Physical Activity
ARIC	Include	Not Include	Include	Include
FHS	Not Include	Include	Include	Include
HyperGEN	Not Include	Include	Include	Not Include
NEO	Include	Not Include	Include	Not Include

Table 3. Setting of Scenarios 1-4 using a mixture of results obtained from cohort-specific genome-wide association analyses conducted with "Complete" LRS and "Partially Missing" LRS, and the inclusion of association results in the meta-analysis using Na ïve, Safe, Complete and Moderator Approaches.

Scenario	**Strata**	**Cohort**			
		ARIC	**FHS**	**HyperGEN**	**NEO**
Scenario 1	Unexposed	Complete[N,S,C,M]	Partially Missing[N,M]	Partially Missing[N,M]	Partially Missing[N,M]
	Exposed	Complete[N,S,C,M]	Partially Missing[N,S,M]	Partially Missing[N,S,M]	Partially Missing[N,S,M]
Scenario 2	Unexposed	Partially Missing[N,M]	Complete[N,S,C,M]	Partially Missing[N,M]	Partially Missing[N,M]
	Exposed	Partially Missing[N,S,M]	Complete[N,S,C,M]	Partially Missing[N,S,M]	Partially Missing[N,S,M]
Scenario 3	Unexposed	Complete[N,S,C,M]	Partially Missing[N,M]	Partially Missing[N,M]	Complete[N,S,C,M]
	Exposed	Complete[N,S,C,M]	Partially Missing[N,S,M]	Partially Missing[N,S,M]	Complete[N,S,C,M]
Scenario 4	Unexposed	Complete[N,S,C,M]	Partially Missing[N,M]	Complete[N,S,C,M]	Complete[N,S,C,M]
	Exposed	Complete[N,S,C,M]	Partially Missing[N,S,M]	Complete[N,S,C,M]	Complete[N,S,C,M]

[N]: included in the meta-analysis using Na ïve Approach; [S]: included in the meta-analysis using Safe Approach; [C]: included in the meta-analysis using Complete Approach; [M]: included in the meta-analysis using Moderator Approach.

Figure 1. Scatterplots of comparison of four approaches to "benchmark" results in terms of –$\log_{10}$ (p-value). Each plot shows SNPs with p-value<10^{-6} for any of the two approaches being compared in the plot. SNPs reaching genome-wide significant (p-value<5×10^{-8}) in "benchmark" results are marked as triangle.

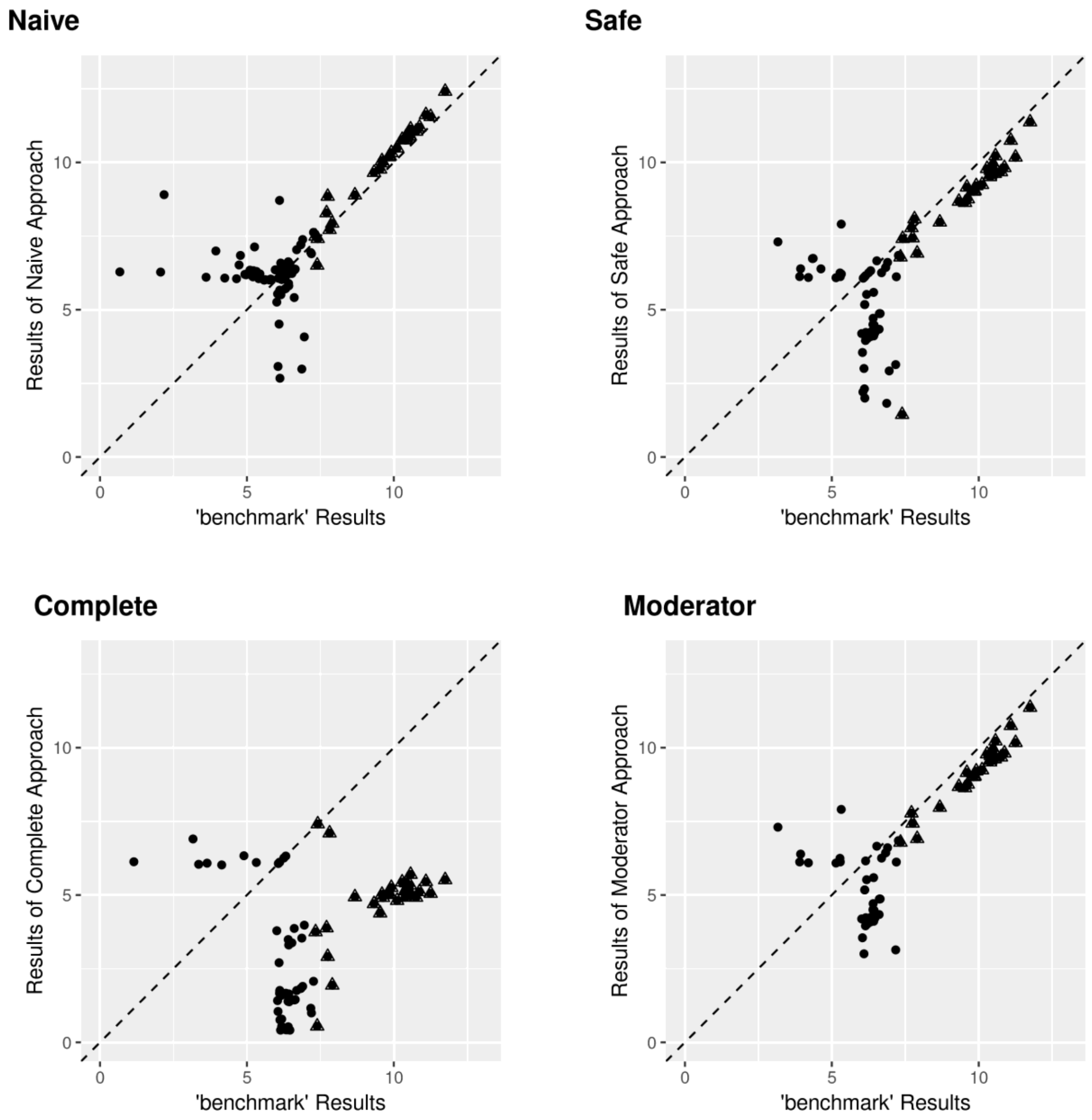


Figure 2: Q-Q plot of different approaches compared to "benchmark" results for Scenario 1 (Scenario 1: use "complete" results from ARIC, and "partially missing" results from HyperGEN, FHS and NEO). $\lambda_{\text{Naive}} = 1.004$, $\lambda_{\text{Safe}} = \lambda_{\text{Moderator}} = 0.985$, $\lambda_{\text{Complete}} = 0.972$.

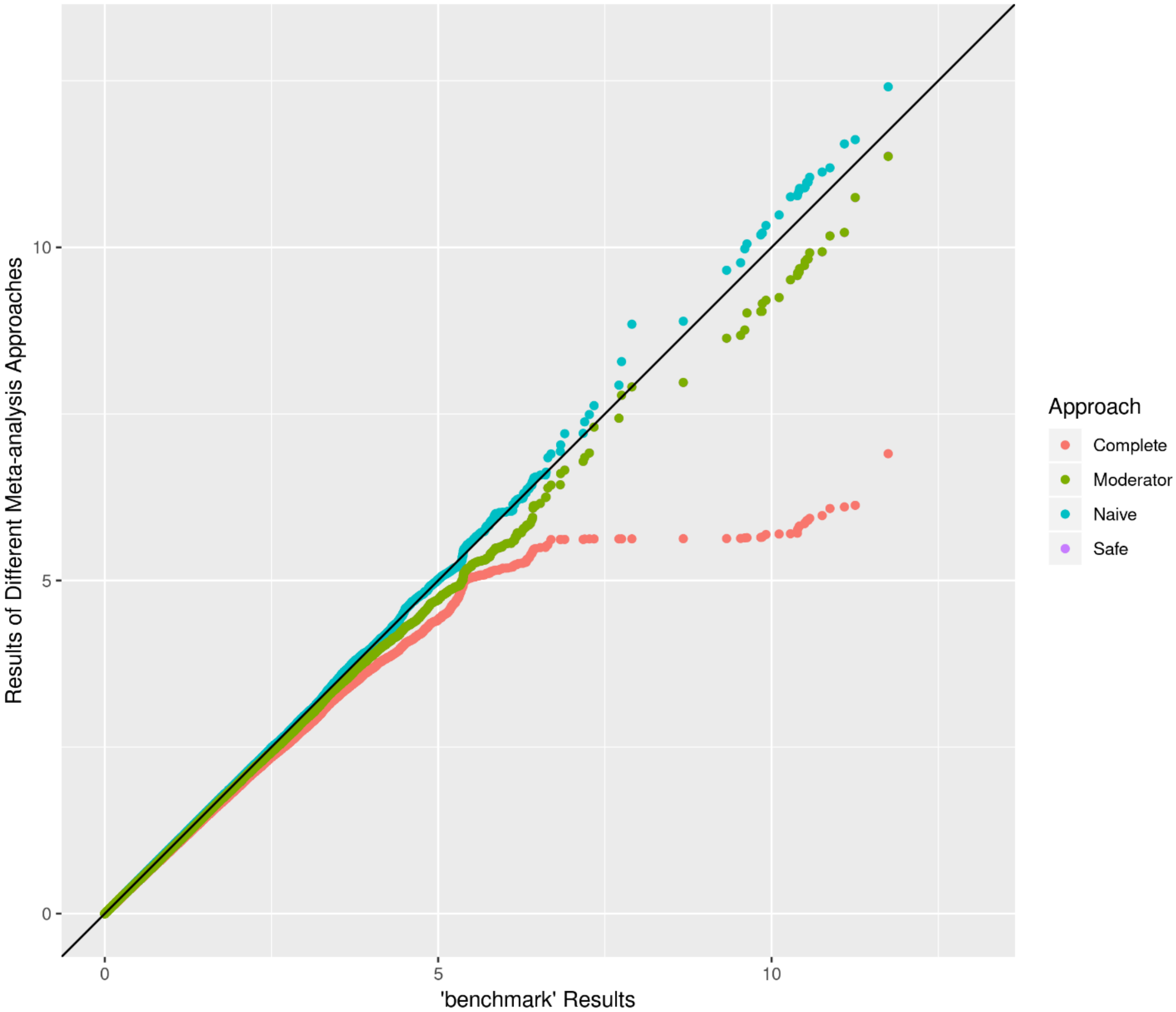


Figure 3. Scatterplots of comparison between four approaches in terms of $-\log_{10}$ (p-value). Each plot shows SNPs with p-value<10^{-6} for any of the two approaches being compared in the plot.

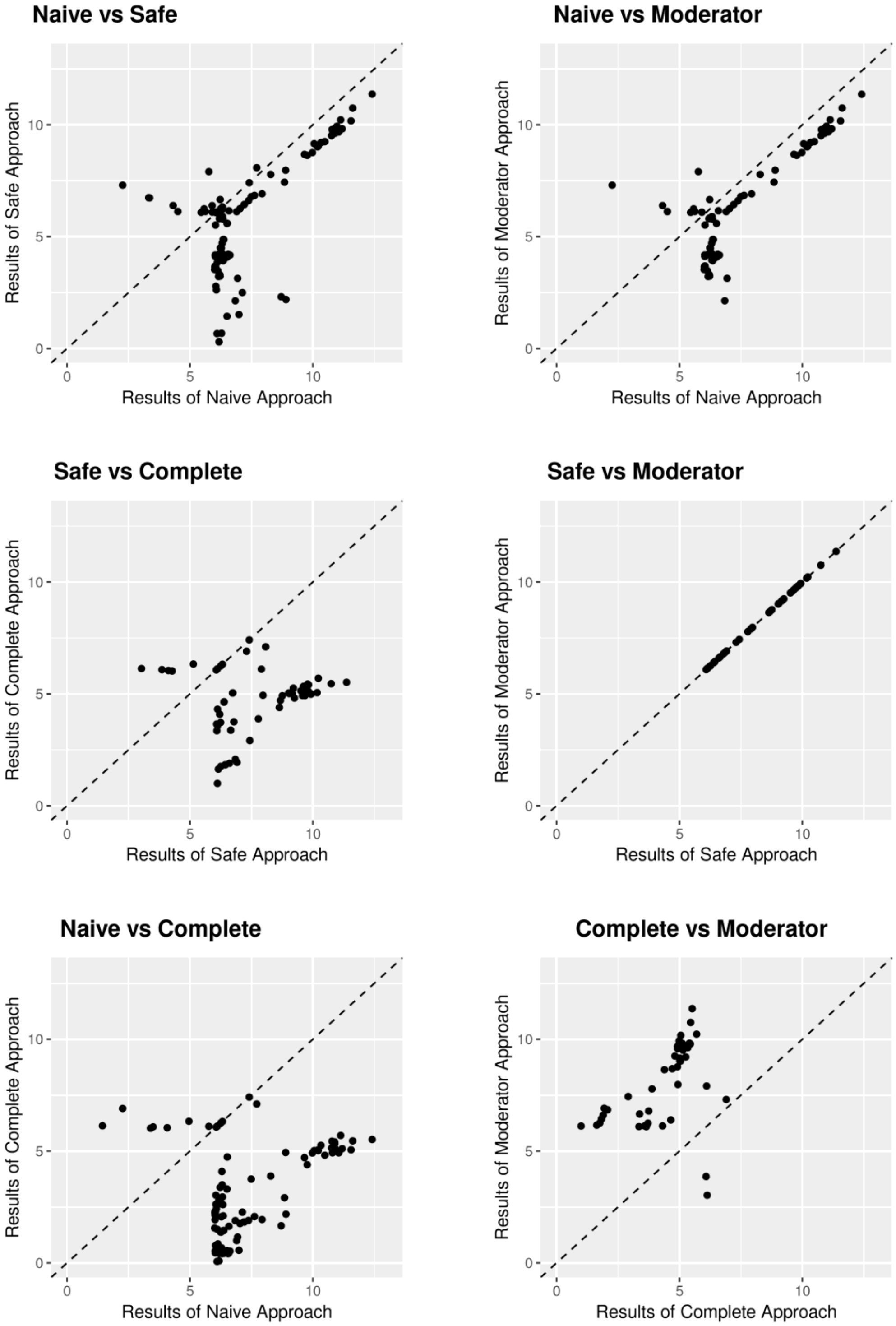
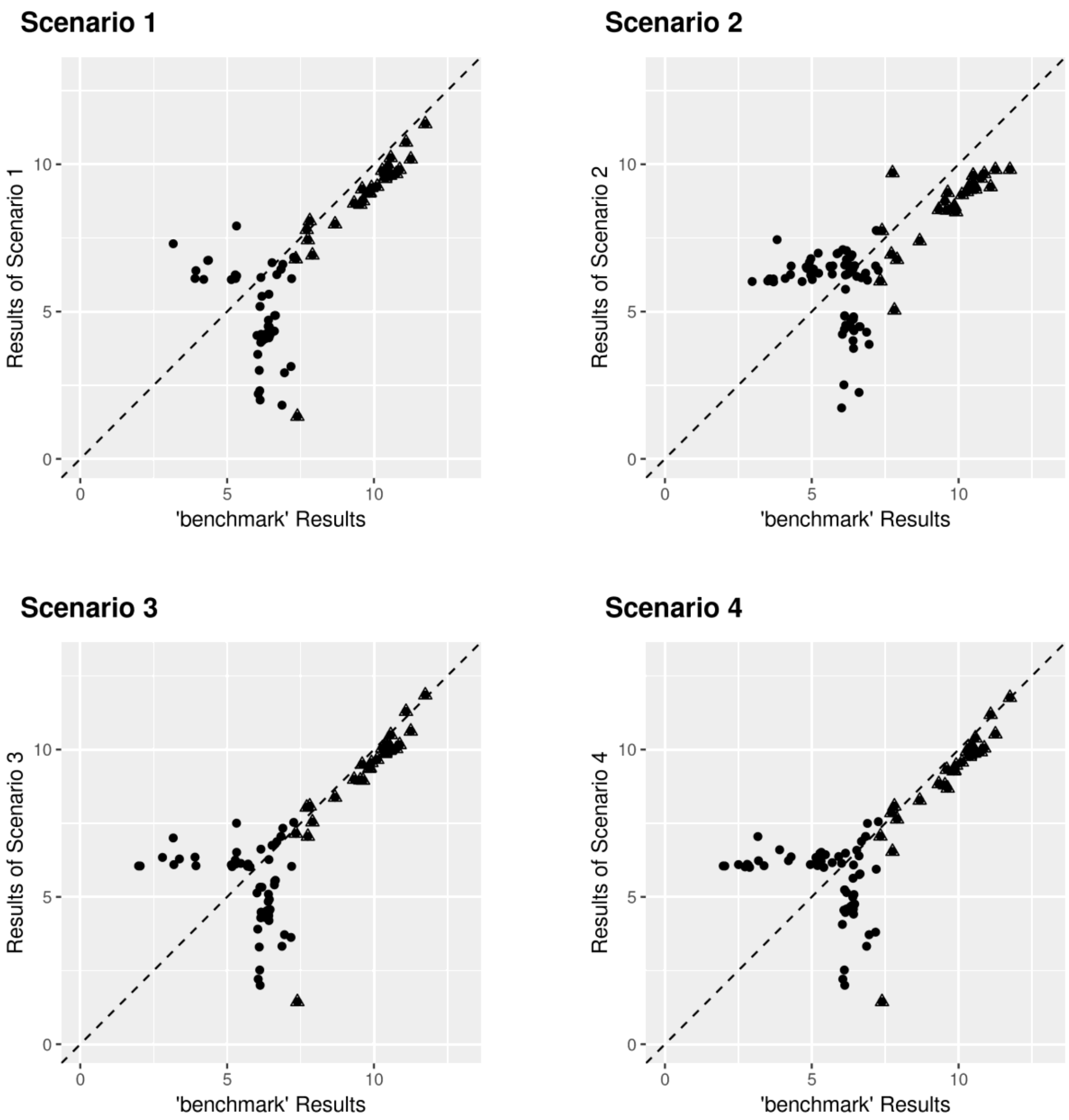


Figure 4. Scatterplots of comparison of four scenarios to “benchmark” results in terms of $-\log_{10}$ (p-value) for Safe Approach. Each plot shows SNPs with p-value<10^{-6} for any of the two approaches being compared in the plot. SNPs reaching genome-wide significant (p-value<5×10^{-8}) in “benchmark” results are marked as triangle.


Scenario 1
Scenario 2
Scenario 3
Scenario 4
Results of Scenario 1
Results of Scenario 2
Results of Scenario 3
Results of Scenario 4
'benchmark' Results