frontiers in Genetics

METHODS
published: 12 October 2021
doi: 10.3389/fgene.2021.745901

Using Summary Statistics to Model Multiplicative Combinations of Initially Analyzed Phenotypes With a Flexible Choice of Covariates

Jack M. Wolf [1], *Jason Westra* [2] *and Nathan Tintle* [2,3]*

[1]Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN, United States, [2]Department of Mathematics, Computer Science, and Statistics, Dordt University, Sioux Center, IA, United States, [3]Department of Population Health Nursing Science, College of Nursing, University of Illinois Chicago, Chicago, IL, United States

While the promise of electronic medical record and biobank data is large, major questions remain about patient privacy, computational hurdles, and data access. One promising area of recent development is pre-computing non-individually identifiable summary statistics to be made publicly available for exploration and downstream analysis. In this manuscript we demonstrate how to utilize pre-computed linear association statistics between individual genetic variants and phenotypes to infer genetic relationships between products of phenotypes (e.g., ratios; logical combinations of binary phenotypes using "and" and "or") with customized covariate choices. We propose a method to approximate covariate adjusted linear models for products and logical combinations of phenotypes using only pre-computed summary statistics. We evaluate our method's accuracy through several simulation studies and an application modeling ratios of fatty acids using data from the Framingham Heart Study. These studies show consistent ability to recapitulate analysis results performed on individual level data including maintenance of the Type I error rate, power, and effect size estimates. An implementation of this proposed method is available in the publicly available R package `pcsstools`.

Keywords: summary statistics, covariate adjustment, linear models, phenotype, multiplication

Edited by:
Pingzhao Hu,
University of Manitoba, Canada

Reviewed by:
Wenjian Bi,
Peking University, China
Rounak Dey,
Harvard University, United States

****Correspondence:***
Nathan Tintle
ntintle@uic.edu

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *22 July 2021*
Accepted: *23 September 2021*
Published: *12 October 2021*

Citation:
Wolf JM, Westra J and Tintle N (2021) Using Summary Statistics to Model Multiplicative Combinations of Initially Analyzed Phenotypes With a Flexible Choice of Covariates. Front. Genet. 12:745901. doi: 10.3389/fgene.2021.745901

1 INTRODUCTION

Researchers now have readily available access to massive quantities of genotypic and phenotypic data (Cox, 2018; Simell et al., 2019). For example, *via* the Electronic Medical Records and Genomics {eMERGE Network[1], the UKBiobank (Bycroft et al., 2018) other initiatives and repositories [e.g., 23andMe, MGI[2] (Gagliano Taliun et al., 2020), FINRISK, CHOP (Diogo et al., 2018), among others]}, researchers can access a wide variety of phenotypic and genomics data on hundreds of thousands of individuals. However, important questions remain about how to best leverage these repositories. For example, the size of biobank datasets makes it challenging to transfer, store, and analyze data locally. While cloud computing minimizes some of these issues, it brings its own challenges related to cost (storage and computation), transfer, and access. Furthermore, data security

[1] https://www.genome.gov/Funded-Programs-Projects/Electronic-Medical-Records-and-Genomics-Network-eMERGE
[2] http://pheweb.sph.umich.edu/

and privacy issues are of paramount importance throughout all aspects of the data access, storage, and analysis pipeline (Jones et al., 2012; Heatherly, 2016; Simell et al., 2019).

A key innovation in this field is pre-computing non-individually identifiable summary statistics on biobank data and maximizing access to this data (Pasaniuc and Price, 2017). For example, GeneAtlas provides basic summary statistics for simple linear regression models of single nucleotide variants (SNVs) with 1,000s of available phenotypic variables across hundreds of thousands of individuals in the UK Biobank (Canela-Xandri et al., 2018), which also provides access to phenotype-phenotype correlations, single nucleotide polymorphism (SNP) minor allele frequencies (MAFs) and Hardy Weinberg Equilibrium (HWE) *p*-values. Likewise, PheWeb[3] is a software toolkit which provides access to UK Biobank and Michigan Genomics Initiative data *via* a series of easy-to-navigate visualization and summary tools (Gagliano Taliun et al., 2020). Others (e.g., The Lee Lab for Statistical Genetics and Data Science[4]) simply provide access to sets of pre-computed summary statistics (PCSS) from large datasets. These resources mitigate many of the privacy and security concerns mentioned above since no individual participant data (IPD) is shared. In addition, the size of these repositories are only fractions of the size of IPD, making transfer and storage of the data much more efficient. Finally, pre-computing these summary statistics alleviates much of the computational burden on researchers who would otherwise have to calculate this information on their own. Despite these advantages, significant limitations currently exist when using these repositories of PCSS.

For example, researchers may want to modify a phenotype with available PCSS to one that is of greater clinical interest or use different sets of covariates than those considered in pre-computed analyses. Recent work is beginning to address these limitations. In two recent papers by our group (Gasdaska et al., 2019; Wolf et al., 2020), we demonstrated how to use standard PCSS (only means, variances, and correlations of all predictors and responses) to calculate the coefficients and standard errors for the linear model for a linear combination of phenotypes with an arbitrary set of covariates. This can then be used to perform Principal Component Analysis (PCA) on a set of phenotypes since principal component scores are just linear combinations with weights derived from the phenotype covariance matrix. Further, we demonstrated that if the phenotype correlation matrix is not available, we can use the correlation of test statistics for each phenotype across all genetic markers in its place with little loss of efficiency. These innovations mean that researchers can, using only PCSS, select the unique set of covariates they wish to adjust for and model a linear combination of phenotypes.

Importantly, these two approaches which require a priori specification of a phenotype of clinical interest, contrast to other recently developed methods which jointly and simultaneously analyze multiple phenotypes (Dutta et al., 2019a,b; Guo and Wu, 2019; Li et al., 2020; Ray and Boehnke, 2018) without an explicit characterization of the relationship between the phenotypes. These joint phenotype tests aim to simultaneously analyze multiple phenotypes while satisfying statistical objectives such as maximizing power under certain conditions. Furthermore, some of these approaches (Ray and Boehnke, 2018; Guo and Wu, 2019) do so using PCSS readily available from existing repositories.

Currently, our group's methods for using PCSS to analyze modified phenotypes with flexible covariate choices are limited to PCA and choosing a phenotype that is a linear combination of the phenotypes for which PCSS are available. Another meaningful way to combine phenotypes is through multiplication. That is, several phenotypes of interest can be viewed as multiplicative combinations of other phenotypes for which PCSS may be available. Examples of note include fatty acid conversion ratios (Kalsbeek et al., 2018) and the body mass index (Justice et al., 2017). Additionally, products of binary phenotypes can be interpreted as logical "and" and "or" statements (e.g., a phenotype $\boldsymbol{y}$ that is defined as "$\boldsymbol{y}_1$ or $\boldsymbol{y}_2$"). Various medical conditions are defined through logical combinations of various phenotypes. For example, coding ischemic strokes based on the union and intersection of various stroke subtypes (von Berg et al., 2020).

In this manuscript, we demonstrate how to analyze modified phenotypes which are multiplicative combinations of an arbitrarily large number of phenotypes for which PCSS are available. We also demonstrate how to flexibly adjust for covariates in these modified phenotype models. Importantly, we also show how the multiplication of phenotypes, when applied to binary phenotypes, allows for logical combination of phenotypes. After presenting a mathematical framework for the method, we validate the method using comprehensive simulations and demonstrate the method on real data from the Framingham Heart Study.

2 METHODS

Consider the m phenotypes $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ where each is an $n \times 1$ vector of measures across n subjects and the $n \times p$ design matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ which consists of variables including genotypic information, covariates, and an intercept column. Moreover, let $\boldsymbol{w}_m = \boldsymbol{y}_1 \boldsymbol{y}_2 \cdots \boldsymbol{y}_m$ denote the pairwise Hadamard product of all m phenotypes for each subject. Our aim is to approximate the coefficients and standard errors of the covariate adjusted linear regression model for the product of m phenotypes: $\boldsymbol{w}_m = \boldsymbol{X}\hat{\boldsymbol{\beta}}+\epsilon$ using only readily available PCSS.

2.1 Assumed Pre-Computed Summary Statistics and Information

We assume knowledge of the following PCSS: the means of every predictor (e.g., SNPs and covariates), the means of every phenotype, and the full variance-covariance matrix of all predictors and phenotypes (i.e., s_{x_j,y_k}, s_{x_j,x_i} and s_{y_k,y_l} for any i, j, k, l where $1 \le i, j \le p$ and $1 \le k, l \le m$). These are all readily

[3] http://pheweb.sph.umich.edu/

[4] https://www.leelabsg.org/resources

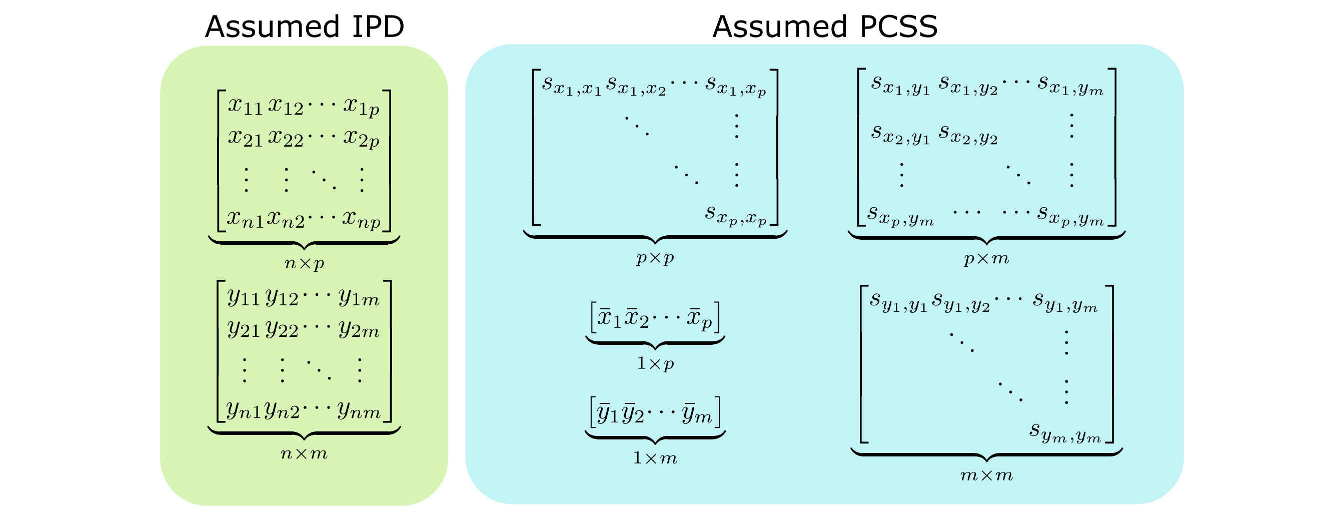

FIGURE 1 | Data assumed when modeling using individual participant data (IPD) and when using pre-computed summary statistics (PCSS) to model a product of m phenotypes ($\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$) as a linear function of p covariates ($\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$). While modeling using IPD requires $n \times (p + m)$ points of data, using PCSS only requires $p^2 + pm + m^2 + p + m$ values, which is far less when n is moderately large compared to p and m. All of these PCSS are readily available in existing PCSS repositories, or can be derived or approximated from other PCSS.

available in standard PCSS repositories. We also assume to know the marginal distribution that each predictor and phenotype follows (e.g., binomial, normal, etc.). **Figure 1** displays the assumed information when modeling *via* both IPD and PCSS.

However, if some summary statistics are unknown, they may be able to be derived or approximated. For example, SNPs distributed in HWE can have their mean and variance approximated through a binomial distribution given the MAF. Furthermore, the covariance of a genetic variant and a non-genetic variable is calculated as the single-marker slope coefficient (for the model with the non-genetic variable as the response and the genetic variant as the predictor) divided by the variance of the genetic variant. Other published papers (Kim et al., 2015; Zhu et al., 2015) have shown that the correlation of two traits can be approximated by the correlation of Z statistics of SNPs not associated with either trait. This approximation method is described in detail in Ray and Boehnke (2018). Two of our previous papers (Gasdaska et al., 2019; Wolf et al., 2020) have demonstrated the accuracy of these three methods through both simulation and real-data applications.

2.2 Linear Regression With Covariates Using Pre-Computed Summary Statistics

Given a response vector $\boldsymbol{w}_m$ and design matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ which includes p variables including SNPs' minor allele counts, covariates, and a possible intercept column, the normal error regression model $\boldsymbol{w}_m = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ where $\epsilon \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ has ordinary least squares estimate for $\boldsymbol{\beta}$: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{w}_m$ with $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$. In a recent paper (Wolf et al., 2020), we demonstrated how to calculate these values using only PCSS using the facts that:

$$\boldsymbol{X}'\boldsymbol{X} = (n-1)S(\boldsymbol{X}) + n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}' \tag{1}$$

$$\boldsymbol{X}'\boldsymbol{w}_m = (n-1)(s_{w_m,x_1}, \ldots, s_{w_m,x_p})' + n\bar{w}_m\bar{\boldsymbol{x}} \tag{2}$$

$$\boldsymbol{w}_m'\boldsymbol{w}_m = (n-1)s^2_{w_m} + n\bar{w}^2_m \tag{3}$$

and

$$\hat{\sigma}^2 = (\boldsymbol{w}_m'\boldsymbol{w}_m - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{w}_m)/(n-p) \tag{4}$$

where $S(\boldsymbol{X})$ is the $p \times p$ variance-covariance matrix of the columns of the design matrix $\boldsymbol{X}$, $\bar{\boldsymbol{x}} = (\bar{x}_1, \ldots, \bar{x}_p)'$ is the $p \times 1$ vector of column means of $\boldsymbol{X}$, $\bar{w}_m$ is the mean of $\boldsymbol{w}_m$, and s_{w_m,x_j} is the sample covariance between $\boldsymbol{w}_m$ and $\boldsymbol{x}_j$.

With these methods in mind and assumed access to standard PCSS, in order to approximate $\hat{\boldsymbol{\beta}}$, and $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ for this covariate adjusted multiple linear regression model, all that remains is to estimate $\bar{w}_m$, $s^2_{w_m}$, and s_{x_j,w_m} for each j. We will first demonstrate how to approximate these values using PCSS when $m = 2$ in **Section 2.3.1** and later show how recursion can be used when $m > 2$ in **Section 2.3.2**.

2.3 Covariance Estimation

2.3.1 Covariance Estimation With the Product of 2 Phenotypes

Let $\boldsymbol{w}_2 = \boldsymbol{y}_1\boldsymbol{y}_2$ be the pairwise Hadamard product of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$. Then, if $\boldsymbol{x}_j$ represents an "intercept" column of the design matrix with all elements unity [i.e., if $\boldsymbol{x}_j = (1,\ldots,1)'$], we set $s_{x_j,w_2} = 0$. Otherwise, we proceed as follows: We first approximate the conditional mean and variance of $\boldsymbol{y}_k$ given $\boldsymbol{x}_j = x$ for $k \in \{1, 2\}$ through a linear regression model as

$$g(y_k|x) = a_{kj} + b_{kj}x \tag{5}$$

and

$$h(y_k|x) = (n-1)(s^2_{y_k} - b_{kj}s_{x_j,y_k})/(n-2) \tag{6}$$

where $b_{kj} = s_{x_j,y_k}/s^2_{x_j}$ and $a_{kj} = \bar{y}_k - b_{kj}\bar{x}_j$. We note that this conditional variance will be constant at any value of x following from the linear regression assumption of homoscedasticity.

Then, we calculate the sample partial correlation of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ controlling for $\boldsymbol{x}_j$:

$$r_{y_1,y_2 \cdot x_j} = \frac{r_{y_1,y_2} - r_{x_j,y_1} r_{x_j,y_2}}{\sqrt{(1 - r^2_{x_j,y_1})(1 - r^2_{x_j,y_2})}}, \tag{7}$$

setting $r_{y_1,y_2 \cdot x_j} = 0$ if either r_{x_j,y_1} or $r_{x_j,y_2} = 1$. As the expectation of the conditional correlation equals the partial correlation under the assumption of a multivariate linear relationship between ($\boldsymbol{y}_1$, $\boldsymbol{y}_2$) and $\boldsymbol{x}_j$ (Baba et al., 2004), we use the partial correlation as an estimate of the conditional correlation of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ at all possible values of $\boldsymbol{x}_j$. So, we approximate the covariance of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ conditional on $\boldsymbol{x}_j$:

$$h(y_1, y_2|x) = r_{y_1,y_2 \cdot x_j}\sqrt{h(y_1|x)h(y_2|x)} \tag{8}$$

These terms let us approximate the conditional mean of $\boldsymbol{w}_2$ at a given value x of $\boldsymbol{x}_j$:

$$g(w_2|x) = g(y_1|x)g(y_2|x) + h(y_1, y_2|x) \tag{9}$$

Then, letting $f_j(x)$ be an assumed probability distribution/mass function for $\boldsymbol{x}_j$ with support $\mathcal{S}_j$ [e.g., if $\boldsymbol{x}_j$ is a vector of minor allele counts with MAF p, letting $f_j(x) = \binom{2}{x}p^x(1-p)^{2-x}$ and $\mathcal{S}_j = \{0, 1, 2\}$) we approximate the sample covariance of $\boldsymbol{x}_j$ and $\boldsymbol{w}_2$:

$$s_{x_j,w_2} \approx \sum_{x \in \mathcal{S}_j} f_j(x)(x - \bar{x}_j)g(w_2|x), \tag{10}$$

swapping the sums for integrals across the support when appropriate.

We calculate the sample mean of $\boldsymbol{w}_2$ as

$$\bar{w}_2 = \bar{y}_1\bar{y}_2 + s_{y_1,y_2}(n-1)/n \tag{11}$$

To approximate the variance, we first approximate the conditional variances of $\boldsymbol{w}_2$ at all levels of $\boldsymbol{x}_j$:

$$\begin{aligned} h(w_2|x) = \quad & h(y_1|x)h(y_2|x) + g(y_1|x)h(y_2|x) + \\ & g(y_2|x)h(y_1|x) + g(w_2|x)h(y_1, y_2|x) \end{aligned} \tag{12}$$

And then approximate the sample variance as:

$$s^2_{w_2} \approx \left(\sum_{x \in \mathcal{S}_j} \left(nf_j(x) - 1\right)h(w_2|x) + nf_j(x)\left(g(w_2|x) - \bar{w}_2\right)^2 \right)/(n-1) \tag{13}$$

once again swapping the sum for an integral across $\mathcal{S}_j$ when appropriate. This approach leads to a different variance estimate for each predictor $\boldsymbol{x}_j$. We treat the median of these estimates across each j as the estimated variance.

Hence, taking the means, variances, and pairwise covariances of $\boldsymbol{x}_j$, $\boldsymbol{y}_1$, and $\boldsymbol{y}_2$ and a distributional assumption about $\boldsymbol{x}_j$, we approximate the covariance of one variable ($\boldsymbol{x}_j$) with the product of the other two ($\boldsymbol{w}_2 = \boldsymbol{y}_1\boldsymbol{y}_2$) as well as the product's mean and variance.

Repeating this algorithm for each predictor $\boldsymbol{x}_j$ and following the linear regression equations presented in **Section 2.2** allows for calculation of covariate adjusted slope coefficients for the multiple regression model $\boldsymbol{w}_2 = \boldsymbol{X}\hat{\boldsymbol{\beta}}+\epsilon$ as well as the standard errors of these slope estimates.

2.3.2 Covariance Estimation With the Product of 3 or More Phenotypes

Regression models for larger products of phenotypes can also be approximated by applying the established method recursively: first estimating the covariance of $\boldsymbol{x}_j$ and $\boldsymbol{w}_2$, then leveraging the covariance of $\boldsymbol{x}_j$ and $\boldsymbol{w}_2$ and $\boldsymbol{x}_j$ and $\boldsymbol{y}_3$ to estimate the covariance of $\boldsymbol{x}_j$ and $\boldsymbol{w}_3$, and so forth. This recursion procedure is described in more detail in the appendix and software to carry it out is discussed in **Section 2.8**.

Let $\boldsymbol{w}_l = \boldsymbol{y}_1\boldsymbol{y}_2\cdots\boldsymbol{y}_l = \boldsymbol{w}_{l-1}\boldsymbol{y}_l$. In order to estimate s_{x_j,w_l} through our established method, we use $\bar{x}_j$, $\bar{y}_l$, $\bar{w}_{l-1}$, $s^2_{x_j}$, $s^2_{y_l}$, $s^2_{w_{l-1}}$, s_{x_j,y_l}, $s_{x_j,w_{l-1}}$, and s_{w_{l-1},y_l} as inputs to the method described in **Section 2.3.1**. That is, replacing $\boldsymbol{y}_1$ with $\boldsymbol{w}_{l-1}$ and $\boldsymbol{y}_2$ with $\boldsymbol{y}_l$. While $\bar{x}_j$, $\bar{y}_l$, $s^2_{x_j}$, $s^2_{y_l}$, and s_{x_j,y_l} are assumed to be known, we must estimate $s_{x_j,w_{l-1}}$ and s_{w_{l-1},y_l}.

Continuation of the recursive process starting at $\boldsymbol{w}_{l-1}$ and working down to $\boldsymbol{w}_2$ will yield an estimate for $s_{x_j,w_{l-1}}$, or eventually the base case of s_{x_j,w_2}.

To approximate s_{w_{l-1},y_l}, we re-express the term as $s_{y_l,w_{l-2}y_{l-1}}$ and approximate the covariance of $\boldsymbol{y}_l$ and $\boldsymbol{w}_{l-2}\boldsymbol{y}_{l-1}$ through the method described in **Section 2.3.1**.

A diagram of the start of this recursion is displayed in **Supplementary Figure S1**.

This recursive estimation is impacted by the order in which the phenotypes are multiplied. So, any set of more than two phenotypes will render $m!/2$ possible ways to estimate the regression model through this method. Hence, we approximate the covariances and means using all permutations of $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ unique up to the order of the first two terms as the order of our phenotypes, and take the median estimate of each term across all permutations as its predicted value.

2.4 Binary Phenotypes

Binary phenotypes present both new challenges to estimation and the opportunity to express logical combinations of phenotypes through products.

2.4.1 Changes to Estimations

Our proposed method leverages several assumptions of the standard linear regression model in its approximations (namely those of homoscedasticity and linearity). As binary phenotypes notoriously violate both of these assumptions, we slightly adjust our approximation to better reflect this situation.

The covariance of two binary phenotypes is estimated using the same general framework as developed in **Section 2.3.1**. The only changes are to mean the variance estimates. All estimated means (conditional and otherwise) are initially estimated as proposed in **Section 2.3.1** and then restricted to the open interval (0, 1) (e.g., letting $g(y_k|x) = a_{kj} + b_{kj}x$ if $0 < a_{kj} + b_{kj}x < 1$, ϵ if $a_{kj} + b_{kj}x \leq 0$, and $1 - \epsilon$ if $1 \leq a_{kj} + b_{kj}x$ for some small $\epsilon > 0$).

Instead of estimating a phenotype's conditional variance from a linear model's residual variance, we estimate it as

$$h(y_k|x) = g(y_k|x)(1 - g(y_k|x)) \tag{14}$$

Further, we calculate the product's sample variance as

$$s^2_{w_2} = \bar{w}_2(1 - \bar{w}_2)n/(n-1) \tag{15}$$

2.4.2 Products as Logical Combinations

Binary phenotypes are of particular importance because their products can be interpreted as logical combinations.

We can represent the logical conjunction $\boldsymbol{y}_1 \wedge \boldsymbol{y}_2$ (read as "$\boldsymbol{y}_1$ and $\boldsymbol{y}_2$") as the product $\boldsymbol{y}_1\boldsymbol{y}_2$. Likewise, we express the logical disjunction $\boldsymbol{y}_1 \vee \boldsymbol{y}_2$ ("$\boldsymbol{y}_1$ or $\boldsymbol{y}_2$") as $\mathbf{1}_n - ((\mathbf{1}_n - \boldsymbol{y}_1)(\mathbf{1}_n - \boldsymbol{y}_2))$.

By framing both disjunctions and conjunctions in terms of phenotype multiplication, we can apply our established methods to approximate the covariances of these combinations with predictors and ultimately estimate linear models for these logical combinations.

While the case of the conjunction is a trivial application of the above methods of multiplying phenotypes, we will briefly describe how to model the disjunction. To do so, we consider the modified phenotypes $\boldsymbol{y}_1' = \mathbf{1}_n - \boldsymbol{y}_1$ and $\boldsymbol{y}_2' = \mathbf{1}_n - \boldsymbol{y}_2$ (these represent the statements "not $\boldsymbol{y}_1$" and "not $\boldsymbol{y}_2$"). This gives us $\boldsymbol{y}_1 \vee \boldsymbol{y}_2 = \mathbf{1}_n - \boldsymbol{y}_1{}'\boldsymbol{y}_2{}'$. Then, $\bar{y}_l' = 1 - \bar{y}_l$, $s_{x_j,y_l'} = -s_{x_j,y_l}$, and $s_{y_k',y_l'} = s_{y_k,y_l}$. If we set $\boldsymbol{w}_2' = \bar{y}_1'\bar{y}_2'$, our method allow us to estimate $s_{x_j,w_2'}$ for each $\boldsymbol{x}_j$ as well as $\bar{w}_2'$ and $s^2_{w'}$. Leveraging these estimates, $s_{x_j,w_2} = -s_{x_j,w_2'}$, $\bar{w}_2 = 1 - \bar{w}_2'$, and $s^2_{w_2} = s^2_{w_2'}$, where $\boldsymbol{w}_2$ is equivalent to the disjunction $\boldsymbol{y}_1 \vee \boldsymbol{y}_2$. Using these terms as inputs for the framework presented in **Section 2.2** allow for coefficient and standard error estimation for the linear model $\boldsymbol{y}_1 \vee \boldsymbol{y}_2 = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \epsilon$.

2.5 Simulation Studies

2.5.1 Simulation 1: Type I Error Maintenance

To verify that our linear model with PCSS approach appropriately maintained the Type I error rate at a variety of α thresholds, we carried out a simulation under the null hypothesis that the predictor variant has no linear association with any of the phenotypes of interest. This null hypothesis represents a reasonable subset of the exact null hypothesis which is that the *product* of phenotypes has no linear relationship with the predictor. We carried out this simulation with varying sample size, MAF, phenotype means, phenotype correlations, and for continuous phenotypes, phenotype variances, for products of two binary phenotypes, two continuous phenotypes, and three continuous phenotypes. When simulating continuous phenotypes we assumed that $Y_{ik} \sim N(\mu_k, \sigma_k^2)$ with Cor(Y_{ik}, Y_{il}) = ρ_{kl} while when simulating binary phenotypes we let $Y_{ik} \sim$ Bernoulli(μ_k) (again with Cor(Y_{ik}, Y_{il}) = ρ_{kl}). Simulation parameters (e.g., n, μ_1, ρ_{12}) were randomly sampled from various distributions (full details are available in **Supplementary Table S1**). We carried out 10^8 simulations for each collection of continuous phenotypes and 10^7 simulations for the case of binary phenotypes.

2.5.2 Simulation 2: Comparisons to IPD Models

To evaluate our method's ability to replicate the results of covariate adjusted linear models fit to IPD, we carried out three 2^k factorial simulations—one for the product of two binary phenotypes, one for the product of two positive continuous phenotypes, and one for the product of three positive continuous phenotypes. We carried out 1,000 simulations at each possible combination of parameters. In each simulation, we modeled the phenotype product as a function of a SNP and binary covariate. For the simulations with only two phenotypes, we also included a continuous covariate in our models.

In all simulations, we simulated n subjects' SNP minor allele counts $\boldsymbol{x}_1$ at HWE with varying MAF. We simulated a binary covariate $\boldsymbol{x}_2 \sim$ Bernoulli(logit$^{-1}\alpha_2\boldsymbol{x}_1$). When generating sets of two phenotypes we also generated a continuous covariate $\boldsymbol{x}_3$ from a linear model with $\boldsymbol{x}_1$ such that $\bar{x}_3 = 0$, $s^2_{x_3} = 1$, and $r_{x_1,x_3} = \alpha_3$. This resulted in a SNP with two covariates (p = 3) in our two phenotype simulations, and a SNP with one covariate (p = 2) in our three phenotype simulation.

We generated individual phenotype measures through the model

$$u(y_{ik}) = \beta_{k0} + \sum_{j=1}^{p} x_{ij}\beta_{kj} + \epsilon_{ik}$$

where $u(y_{ik}) = y_{ik}$ for continuous phenotypes, $u(y_{ik})$ = logit(Pr(Y_{ik} = 1)) for binary phenotypes, and ϵ_i' follows a multivariate normal distribution with $\boldsymbol{\mu} = \boldsymbol{0}$ and $\boldsymbol{\Sigma}_{(i,j)} = \sigma_i\sigma_j\rho_{ij}$. Parameter values were selected such that, under optimal settings, empirical power was roughly 80–90% at a significance threshold of 10^{-8}. Full details of simulation parameters are available in **Supplementary Table S2**.

In each simulation, we estimated coefficients, standard errors, t statistics, and two-sided p-values for the null hypothesis that there was no relationship between the product of phenotypes and the SNP ($\boldsymbol{x}_1$) after adjusting for covariates both using IPD and using PCSS.

Additionally, when simulating two binary phenotypes we fit covariate-adjusted logistic regression models for the logged odds that $y_{1i}y_{2i} = 1$ using IPD and returned the relevant two-sided p-value to compare the results of the linear model fit using PCSS to the correctly specified logistic model.

2.6 Real Data Application

Fatty acids are of broad importance for a wide range of cardiometabolic traits (Imamura et al., 2020) with ratios of fatty acids are often used as a proxy for conversion efficiency. Previous genome wide association studies have explored the genetic architecture of fatty acids and their ratios (Kalsbeek et al., 2018; Lemaitre et al., 2011; Tintle et al., 2015, 2020).

We modeled 12 fatty acid ratios using both IPD and PCSS using data from the Framingham Heart Study's Generation-3 and Offspring cohorts downloaded from dbGaP (Mailman et al., 2007). The specific ratios can be found in the first column of **Table 3**. **Supplementary Table S3** lists all fatty acids used in at least one of the ratios alongside their abbreviations.

Quality control measures included setting Mendelian inconsistencies as missing and excluding SNPs with HWE $p <$ 0.00001, MAF <0.05, or missing values for over 10% of subjects. We excluded individuals missing over 10% of their genetic data after initial quality control and then took a subset of unrelated participants. After quality control we were left with 362,330 SNPs over 1,455 individuals (657 from the Offspring cohort and 888 from the Generation-3 cohort).

In addition to the standard PCSS described in **Section 2.1**, we assumed access to pre-computed means and variances of the reciprocal of each fatty acid as well as the correlation between any fatty acid reciprocal and any other fatty acid, covariate, or SNP to model these ratios using PCSS.

We analyzed each fatty acid ratio through the linear model: Ratio ~ SNP + age + sex for each SNP in our sample using both IPD and PCSS and tested each SNP for statistical significance with the Bonferroni adjusted threshold $\alpha = 1.37 \times 10^{-7}$.

2.7 Statistical Analysis

2.7.1 Simulation 1

To analyze the results of our Type I Error simulations we calculated the empirical Type I Error rate when approximating linear models using PCSS at each specified significance threshold.

2.7.2 Simulation 2

For each of the three 2^k factorial simulations, we assessed the PCSS model's accuracy relative to its IPD counterpart.

We calculated the bias and mean squared error when estimating the SNP's slope coefficient, standard error, and absolute value test statistic. We also modeled errors estimating the slope coefficient, standard error, and test statistic through multiple linear regression models with logical indicators for each of the k parameter settings as predictors, testing at the Bonferroni adjusted significance threshold of $0.05/k$ to determine which simulation parameters affected our method's accuracy.

We compared test decisions regarding the significance of the SNP when modeling the phenotype product after adjusting for covariates at significance thresholds $10^{-1}, 10^{-2}, \ldots, 10^{-8}$. When analyzing binary phenotypes we also compared test decisions between the linear model fit using PCSS and the logistic regression model fit on IPD to demonstrate the robustness of linear models to model binary outcomes.

TABLE 1 | Simulation studies of Type I error estimates when testing the linear association between a single SNP and a product of phenotypes using pre-computed summary statistics at significance thresholds: $\alpha = 0.05$, 0.001, 10^{-5}, and 10^{-6}. Each entry represents the proportion of p-values smaller than α when modeling the linear relation between a SNP and a product of phenotypes.

	Nominal α			
Phenotypes	5.0E-02	1.0E-03	1.0E-05	1.0E-06
2 Continuous	3.88E-02	6.65E-04	5.56E-06	4.40E-07
2 Binary	2.39E-02	2.06E-04	8.00E-07	1.00E-07
3 Continuous	2.70E-02	3.81E-04	2.91E-06	3.40E-07

2.7.3 Real Data Application

We measured our overall bias and mean squared error in slope, standard error, and absolute value test statistic estimates for each outcome modeled. We recorded test decisions for both the IPD and PCSS models and whether the two results agreed or disagreed regarding the statistical significance of each SNP. When one approach found a SNP to be significant and the other did not, we noted if the non-significant result was "borderline" significant ($\alpha \le p < 10\alpha$).

2.8 Software

Software to perform these model approximations as well as those developed in Wolf et al. (2020) is available on CRAN in the R package pcsstools[5].

3 RESULTS

3.1 Simulation Studies

3.1.1 Simulation 1

Empirical Type I error rates when using PCSS are displayed in **Table 1**. In all simulations, the approach's empirical Type I error rate was below the tested significance threshold.

3.1.2 Simulation 2

The PCSS method's errors when approximating slope coefficients, their standard errors, and test statistics are available in **Table 2**. When aggregated over all simulation settings we observe slight positive bias both when estimating the slope and the absolute value of the t test statistic for each collection of phenotypes. The magnitude of the mean test statistic error is comparable across all three simulations. **Figure 2** displays our PCSS method's approximated slope coefficients compared to slope coefficients calculated using IPD for the SNP while modeling the phenotype product and adjusting for covariates. Similar graphical comparisons of standard error and test statistic estimates are available in **Supplementary Figures S2, S3**.

When modeling estimation errors for two continuous phenotypes through a linear regression model with indicator variables for all of the simulation settings ($k = 12$, $n = 2^k \times 10^3$), our model for the slope error found all settings except the residual phenotype variances, σ_k^2, to be significantly associated with the PCSS model's slope estimate's error at the adjusted significance threshold $0.05/k$. All settings had significant associations with our error when estimating the standard error of the slope coefficient, or the test statistic. In the case of two binary phenotypes ($k = 14$, $n = 2^k \times 10^3$), we found all settings to have significant associations with the error in slope, standard error, and test statistic estimates. For three continuous phenotypes ($k = 13$, $n = 2^k \times 10^3$), we also found all settings to have significant associations with the error when predicting the slope coefficient, its standard error, and its test statistic.

Figure 3 shows comparisons of estimated and calculated p-values for a two-sided t test under the null hypothesis that the

[5] https://cran.r-project.org/package=pcsstools

TABLE 2 | Simulation study approximating a linear model for a product of phenotypes using summary statistics. Summaries of errors when approximating slopes, slope standard errors, and the absolute value of the *t*-statistics for a SNP while adjusting for covariates when using pre-computed summary statistics (PCSS) compared to values obtained when calculating these statistics using individual participant data (IPD).

Phenotypes	β			SE(β)		|*t*|	
	IPD Mean	**Bias**	**MSE**	**Bias**	**MSE**	**Bias**	**MSE**
2 Continuous	6.09E-03	3.72E-04	1.43E-05	2.39E-06	6.58E-11	2.33E-03	4.13E-01
2 Binary	4.13E-01	2.56E-02	6.74E-03	6.58E-05	5.50E-07	1.65E-01	2.16E-01
3 Continuous	4.82E+00	3.33E-02	3.79E-01	-3.63E-02	5.82E-04	5.71E-02	1.06E-01

SNP had no linear association with the phenotype product after adjusting for covariates. (**Figure 3** only includes *p*-values greater than 10^{-15} for the sake of visual clarity; **Supplementary Figure S4** repeats this visualization without any restrictions.) **Figure 4** shows various error rates rate between the IPD and PCSS models' test decisions based on these *p*-values at differing significance thresholds. We see that all PCSS models overall disagreement rates to their IPD companions decrease as the significance threshold becomes more stringent. Likewise, when the IPD model rejected the null hypothesis, the PCSS model rarely failed to reject with error rates at most 13% which again decreased as the significance threshold decreased. When the IPD model failed to reject the null hypothesis, the PCSS approach's conditional error rate varied by the model's response. When modeling the product of two continuous or binary phenotypes, the error rate stayed relatively constant across all thresholds at around 3 and 15%, respectively. But, when modeling the product of three continuous phenotypes, the error rate increased as the significance threshold became more strict. Lastly, when compared to the test decisions of a covariate adjusted logistic regression model, our PCSS approximation of the related linear model tends to reach

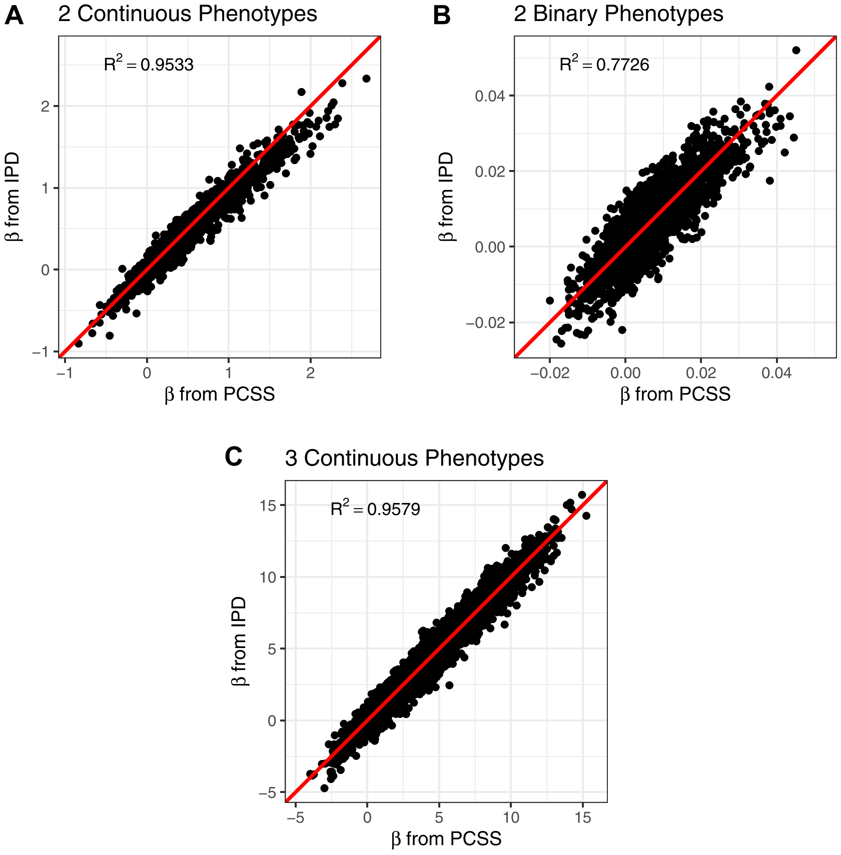

FIGURE 2 | Comparison of slope coefficients from a simulation study approximating a covariate adjusted linear model for a product of phenotypes using pre-computed summary statistics (PCSS) and individual participant data (IPD). **(A)** Modeling the product of two continuous phenotypes while adjusting for a binary and a continuous covariate. **(B)** Modeling the product of two binary phenotypes while adjusting for a binary and a continuous covariate. **(C)** Modeling the product of three continuous phenotypes while adjusting for a binary covariate.

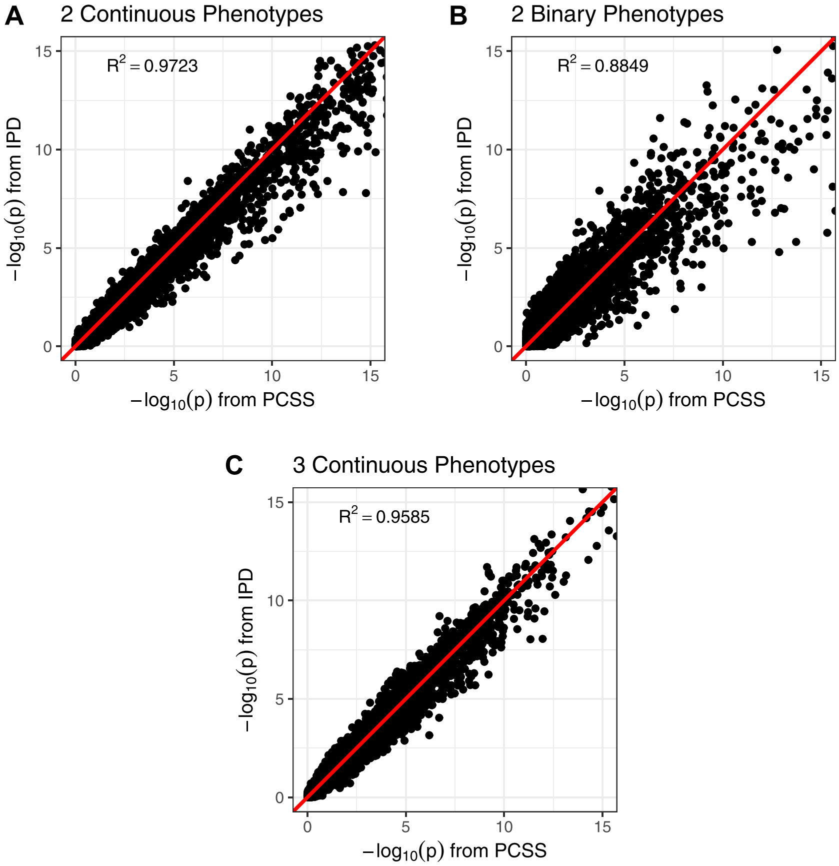

FIGURE 3 | Comparison of *p*-values from a simulation study approximating a covariate adjusted linear model for a product of phenotypes using pre-computed summary statistics (PCSS) and individual participant data (IPD). Two-sided *p*-values were computed for the null hypothesis that the SNP had no linear effect on the phenotype product while adjusting for covariates. All plots are restricted to the range (0, 15); unrestricted plots are available in the supplementary materials. **(A)** Modeling the product of two continuous phenotypes while adjusting for a binary and a continuous covariate. **(B)** Modeling the product of two binary phenotypes while adjusting for a binary and a continuous covariate. **(C)** Modeling the product of three continuous phenotypes while adjusting for a binary covariate.

the same conclusions, with a moderate conservative tendency, especially at more strict significance thresholds.

3.2 Real Data Application

The bias and mean squared error of the PCSS model's approximation to the IPD model's slope, standard error, and absolute value test statistic for each fatty acid ratio are displayed in **Table 3**. Our mean slope error was -2.93×10^{-3} (Mean Squared Error 0.114) while the mean slope estimate when using IPD was -1.3×10^{-3}, demonstrating a slight bias towards zero. However, the standard error estimates under the PCSS model were on average 4.80×10^{-2} lower than their respective estimates under the IPD model and absolute value tests statistics tended to be 1.79×10^{-2} higher than their IPD counterparts indicating an overall minor anti-conservative bias.

Table 4 summarizes the number of SNPs found significant when modeling using both IPD and PCSS across all 12 × 362330 models. Of the ten SNPs for which IPD and PCSS models disagreed, nine occurred when one approach found a SNP to have a significant effect while the other found it to have a borderline significant effect ($\alpha \leq p < 10\alpha$).

4 DISCUSSION

We have developed a method that approximates the covariance of products of phenotypes with other variables using only bivariate and univariate pre-computed summary statistics (PCSS). We then demonstrated how this covariance estimation can be used to approximate linear models for products of phenotypes, how these can model logical "and" and "or" statements and how these models can include researchers choice of covariates. We demonstrated our approximation method's accuracy relative to models fit on individual participant data through multiple simulations and applications to real genetic data.

The approximations showed good performance overall. In a wide variety of simulations, the Type I error was maintained, the bias in point estimates relative to models fit with IPD was minimal, and hypothesis tests almost always reached the same conclusion as would be obtained with IPD. Application of our method to real data from the Framingham Heart Study also showed good performance on similar metrics. In general, we have tried to formulate this PCSS method to only rely on commonly available or easily estimated PCSS. However, in our application we assumed that we had the PCSS for reciprocals of

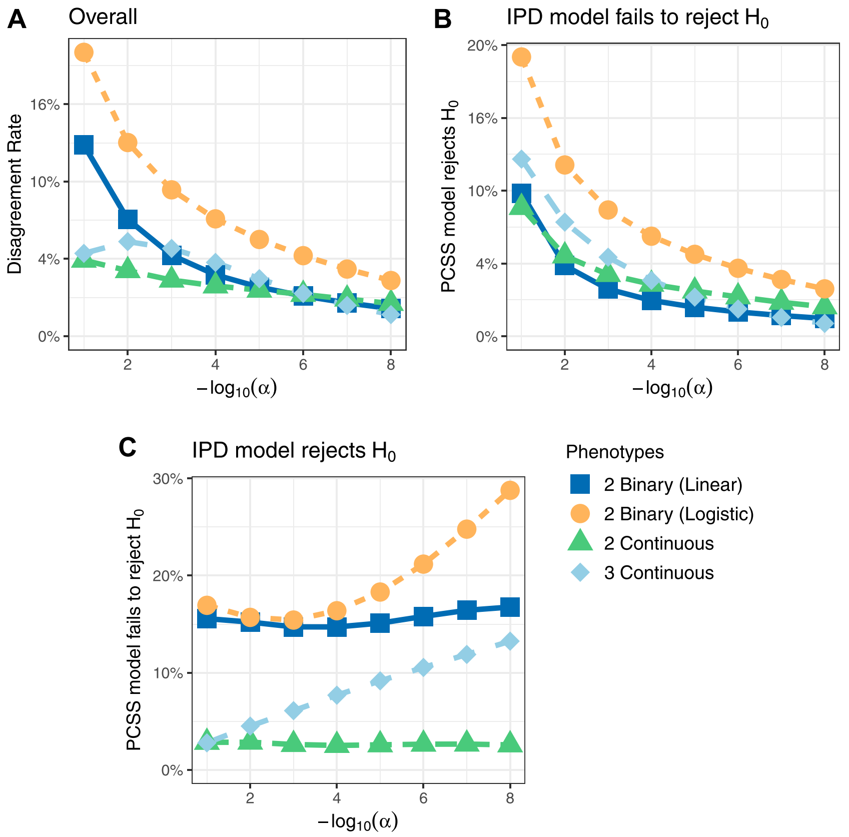

FIGURE 4 | Simulation studies' test decision disagreement rates evaluating the significance of a SNP in a linear model for a product of phenotypes while adjusting for covariates using individualized participant data (IPD) and pre-computed summary statistics (PCSS) at various significance thresholds (α). Comparisons were also made between a logistic regression model fit using IPD on the product of two binary phenotypes and the PCSS model approximating the linear relationship. **(A)** Percentage of times the PCSS and IPD models' test decisions disagreed across all simulations. **(B)** Percentage of times the PCSS model rejected the null hypothesis given that the IPD model failed to reject the null hypothesis. **(C)** Percentage of times the PCSS model failed to reject the null hypothesis given that the IPD model rejected the null hypothesis.

TABLE 3 | Summary of errors when approximating the linear model: FA Ratio ~ snp + age + sex using pre-computed summary statistics (PCSS) compared to values obtained when calculating these statistics using individual participant data (IPD). Each fatty acid ratio was modeled across 362,330 SNPs from 1,455 subjects in the Framing Heart Study's Offspring and Generation-3 cohorts.

Ratio	β			SE(β)		$\|t\|$	
	IPD Mean	Bias	MSE	Bias	MSE	Bias	MSE
PA:POA	5.58E-02	1.49E-02	4.18E-02	−4.80E-02	2.08E-04	4.38E-03	1.45E-02
PA:SA	5.83E-05	−2.13E-05	3.09E-07	1.43E-04	1.87E-09	4.67E-03	8.23E-03
POA:OA	−2.08E-05	−2.10E-06	8.08E-09	−2.69E-05	4.69E-11	5.92E-03	1.70E-02
SA:OA	−8.15E-05	−1.19E-05	3.62E-07	−3.33E-05	3.36E-10	6.86E-03	8.21E-03
LA:GLA	−7.35E-02	2.03E-02	1.32E+00	−5.10E-01	1.95E-02	3.11E-02	3.17E-02
LA:DGLA	2.04E-03	1.99E-04	3.58E-04	−1.84E-03	3.12E-07	2.33E-02	4.08E-02
GLA:DGLA	4.58E-05	1.04E-05	2.64E-07	−7.62E-06	1.13E-10	3.36E-02	6.50E-02
DGLA:AA	−1.99E-05	−1.54E-06	2.84E-08	3.79E-05	1.09E-10	9.37E-03	1.33E-02
AA:DTA	1.76E-03	−4.20E-04	7.21E-05	−3.76E-03	9.42E-07	9.73E-03	2.38E-02
EPA:DPA_N3	2.29E-04	−1.38E-05	1.31E-06	−4.04E-04	1.02E-08	2.34E-02	4.72E-02
DTA:DPA_N6	−1.49E-03	1.66E-04	7.46E-04	−1.16E-02	8.24E-06	4.71E-02	1.43E-01
DPA_N3:DHA	−4.12E-04	6.03E-05	3.45E-06	−1.52E-04	2.53E-09	1.56E-02	3.75E-02
Overall	−1.30E-03	2.93E-03	1.14E-01	−4.80E-02	2.12E-02	1.79E-02	3.77E-02

fatty acids. This may not always be the case in practice, but may suggest that these PCSS may be important to pre-compute to assist downstream analyses of ratios.

Despite these positive results, some limitations of our work are worth noting. First, we used linear regression for a binary response. Previous applications of PCSS have taken this

TABLE 4 | Summary of test decisions for a real data application calculating the linear model Fatty Acid Ratio ~ snp + age + sex using individual participant data (IPD) and pre-computed summary statistics (PCSS) across 362,330 SNPs from 1,455 subjects in the Framing Heart Study's Offspring and Generation-3 cohorts. Significance threshold of $\alpha = 1.37 \times 10^{-7}$.

Ratio	IPD significant	PCSS significant	Both significant
PA:POA	0	0	0
PA:SA	0	0	0
POA:OA	0	0	0
SA:OA	6	9	6
LA:GLA	5	2	2
LA:DGLA	9	10	9
GLA:DGLA	8	8	8
DGLA:AA	18	19	18
AA:DTA	0	0	0
EPA:DPA_N3	0	1	0
DTA:DPA_N6	5	4	4
DPA_N3:DHA	11	11	11
Overall	62	64	58

approach (Canela-Xandri et al., 2018), and it is generally robust; however, this approach is less precise than when the underlying relationship is truly linear. While some foundations for a logistic modelling approach were recently proposed by Wu et al. (2021), further work is needed to develop a comprehensive model for logistic regression using PCSS. We also note that our method makes assumptions about the fit of the linear model to the data. While these assumptions are the same as in the corresponding analysis of IPD data (e.g., true underlying linear relationship between $\boldsymbol{y}$ and $\boldsymbol{x}$), these assumptions may be more acutely important in our PCSS method.

Second, while our simulation study was comprehensive and we demonstrated our method on real data we note that further testing on simulated and real data is encouraged to explore special cases not considered here. Situations that suggest further testing and methodological developments include modeling linear combinations of products and adjusting for clustered/family data. One scenario of particular importance that warrants further investigation is missing data resulting in PCSS being computed on different subsets of the full data. While our real data analysis showed good performance in a setting with missing genotype data, further investigations should be performed to address the robustness of this method (along with other methods based around PCSS) when dealing with missing phenotype information. Similarly, researchers may be limited in their ability to have additional *a priori* exclusion criteria applied to their analysis or obtain PCSS calculated using the same criteria for each phenotype. There are, however, two options. Either obtain summary statistics on the subgroup of interest or ignore the exclusion criteria by including the group in the "controls" in an analysis on a dichotomous outcome. More robust options are in development (e.g., multinomial regression, or methods that use summary statistics alone to allow researchers to apply exclusion criteria). Relatedly, while our method exhibited fair performance when modeling logical combinations of binary phenotypes with low case-control ratios (see **Supplementary Table S4**), it would benefit from further and more thorough work to assess its robustness.

Third, when estimating the variance of a product of a set of continuous phenotypes we estimated this term as a function of each covariate and took the median all these estimates as the approximated value. While this approach works well in practice, it may be possible to utilize the joint distribution of the covariates to estimate the covariance of these estimations and derive a more optimal estimation. Relatedly, our simulations showed that when multiplying binary phenotypes that exhibit high negative correlation and when multiplying phenotypes that take on negative values care should be taken. Finally and relatedly is the issue of potential compounding of errors when the method is applied to products of m phenotypes (where m is large). Although there are meaningful combined phenotypes that consist of five or more phenotypes (upon which this method should be used cautiously), we note that many combined phenotypes [e.g., BMI, ratios of biomarkers, cardiovascular disease (defined as either coronary heart disease or stroke)] are combinations of four or fewer phenotypes. Additional simulation studies and methodological improvements are needed and caution should be exhibited when applying our method in these cases.

Fourth and finally, this method does not support PCSS that describe score tests (where a null model with non-genetic covariates is first fit and then updated for each genetic variant instead of simultaneously estimating both the genetic and covariate effects) in its current form. While future research can likely expanded this method to work with this data, we note the considerable collection of PCSS repositories (e.g., PheWeb, GWAS Catalog, GeneAtlas) which do provide the PCSS needed to perform these proposed approximations.

The use of PCSS provides numerous advantages over IPD data including computational efficiency and reduced concerns about data privacy. However, substantially improved and flexible methods are needed in order to fully leverage PCSS in customized downstream analyses. Our method allows researchers further customization of analyzed phenotypes by opening the door to multiplicative combinations of phenotypes, including logical combinations of binary phenotypes. Approximations used are reasonable, with near perfect maintenance of the Type I error rate and power in most situations. Further work is needed to apply the method to additional datasets and to expand the method to larger classes of combined phenotypes.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/gap/ (accessions phs000007.v29.p10 and phs000342.v20.p13).

AUTHOR CONTRIBUTIONS

JWo and JWe developed the proposed statistical method. JWo performed the simulation analyses. JWe performed the real data analysis. JWo wrote the first draft of the manuscript. JWo and NT wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

The authors of this work were supported by NIH Grant 2R15HG006915-03 and Dordt University.

ACKNOWLEDGMENTS

The authors would like to thank Martha Barnard, Xueting Xia, Nathan Ryder, and Jason Vander Woude for their help with preliminary stages of this project. This manuscript originally appeared online as a preprint on bioRxiv (Wolf et al., 2021).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.745901/full#supplementary-material

REFERENCES

Baba, K., Shibata, R., and Sibuya, M. (2004). Partial Correlation and Conditional Correlation as Measures of Conditional independence. *Aust. New Zealand J. Stat.* 46, 657–664. doi:10.1111/j.1467-842X.2004.00360.x

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., et al. (2018). The UK Biobank Resource with Deep Phenotyping and Genomic Data. *Nature* 562, 203–209. doi:10.1038/s41586-018-0579-z

Canela-Xandri, O., Rawlik, K., and Tenesa, A. (2018). An Atlas of Genetic Associations in UK Biobank. *Nat. Genet.* 50, 1593–1599. doi:10.1038/s41588-018-0248-z

Cox, N. (2018). UK Biobank Shares the Promise of Big Data. *Nature* 562, 194–195. doi:10.1038/d41586-018-06948-3

Diogo, D., Tian, C., Franklin, C. S., Alanne-Kinnunen, M., March, M., Spencer, C. C. A., et al. (2018). Phenome-wide Association Studies across Large Population Cohorts Support Drug Target Validation. *Nat. Commun.* 9, 4285. doi:10.1038/s41467-018-06540-3

Dutta, D., Gagliano Taliun, S. A., Weinstock, J. S., Zawistowski, M., Sidore, C., Fritsche, L. G., et al. (2019a). Meta-MultiSKAT: Multiple Phenotype Meta-Analysis for Region-Based Association Test. *Genet. Epidemiol.* 43, 800–814. doi:10.1002/gepi.22248

Dutta, D., Scott, L., Boehnke, M., and Lee, S. (2019b). Multi-SKAT: General Framework to Test for Rare-Variant Association with Multiple Phenotypes. *Genet. Epidemiol.* 43, 4–23. doi:10.1002/gepi.22156

Gagliano Taliun, S. A., VandeHaar, P., Boughton, A. P., Welch, R. P., Taliun, D., Schmidt, E. M., et al. (2020). Exploring and Visualizing Large-Scale Genetic Associations by Using PheWeb. *Nat. Genet.* 52, 550–552. doi:10.1038/s41588-020-0622-5

Gasdaska, A., Friend, D., Chen, R., Westra, J., Zawistowski, M., Lindsey, W., et al. (2019). Leveraging Summary Statistics to Make Inferences about Complex Phenotypes in Large Biobanks. *Pac. Symp. Biocomputing* 24, 391–402. doi:10.1142/9789813279827_0036

Guo, B., and Wu, B. (2019). Integrate Multiple Traits to Detect Novel Trait–Gene Association Using GWAS Summary Data with an Adaptive Test Approach. *Bioinformatics* 35, 2251–2257. doi:10.1093/bioinformatics/bty961

Heatherly, R. (2016). Privacy and Security within Biobanking: The Role of Information Technology. *J. L. Med. Ethics* 44, 156–160. doi:10.1177/1073110516644206

Imamura, F., Fretts, A. M., Marklund, M., Korat, A. V. A., Yang, W.-S., Lankinen, M., et al. (2020). Fatty Acids in the De Novo Lipogenesis Pathway and Incidence of Type 2 Diabetes: A Pooled Analysis of Prospective Cohort Studies. *PLOS Med.* 17, e1003102. doi:10.1371/journal.pmed.1003102

Jones, E. M., Sheehan, N. A., Masca, N., Wallace, S. E., Murtagh, M. J., and Burton, P. R. (2012). DataSHIELD – Shared Individual-Level Analysis without Sharing the Data: a Biostatistical Perspective. *Norsk Epidemiologi* 21, 1499. doi:10.5324/nje.v21i2

Justice, A. E., Winkler, T. W., Feitosa, M. F., Graff, M., Fisher, V. A., Young, K., et al. (2017). Genome-wide Meta-Analysis of 241,258 Adults Accounting for Smoking Behaviour Identifies Novel Loci for Obesity Traits. *Nat. Commun.* 8, 14977. doi:10.1038/ncomms14977

Kalsbeek, A., Veenstra, J., Westra, J., Disselkoen, C., Koch, K., McKenzie, K. A., et al. (2018). A Genome-wide Association Study of Red-Blood Cell Fatty Acids and Ratios Incorporating Dietary Covariates: Framingham Heart Study Offspring Cohort. *PLoS ONE* 13. doi:10.1371/journal.pone.0194882

Kim, J., Bai, Y., and Pan, W. (2015). An Adaptive Association Test for Multiple Phenotypes with GWAS Summary Statistics. *Genet. Epidemiol.* 39, 651–663. doi:10.1002/gepi.21931

Lemaitre, R. N., Tanaka, T., Tang, W., Manichaikul, A., Foy, M., Kabagambe, E. K., et al. (2011). Genetic Loci Associated with Plasma Phospholipid N-3 Fatty Acids: A Meta-Analysis of Genome-wide Association Studies from the CHARGE Consortium. *PLOS Genet.* 7, e1002193. doi:10.1371/journal.pgen.1002193

Li, X., Zhang, S., and Sha, Q. (2020). Joint Analysis of Multiple Phenotypes Using a Clustering Linear Combination Method Based on Hierarchical Clustering. *Genet. Epidemiol.* 44, 67–78. doi:10.1002/gepi.22263

Mailman, M. D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., et al. (2007). The NCBI dbGaP Database of Genotypes and Phenotypes. *Nat. Genet.* 39, 1181–1186. doi:10.1038/ng1007-1181

Pasaniuc, B., and Price, A. L. (2017). Dissecting the Genetics of Complex Traits Using Summary Association Statistics. *Nat. Rev. Genet.* 18, 117–127. doi:10.1038/nrg.2016.142

Ray, D., and Boehnke, M. (2018). Methods for Meta-Analysis of Multiple Traits Using GWAS Summary Statistics. *Genet. Epidemiol.* 42, 134–145. doi:10.1002/gepi.22105

Simell, B. A., Törnwall, O. M., Hämäläinen, I., Wichmann, H. E., Anton, G., Brennan, P., et al. (2019). Transnational Access to Large Prospective Cohorts in Europe: Current Trends and Unmet Needs. *New Biotechnol.* 49, 98–103. doi:10.1016/j.nbt.2018.10.001

Tintle, N., Bassett, J., Kuo-Liong, C., Forouhi, N., Kupers, L., Lankinen, M., et al. (2020). Circulating Omega-3 Fatty Acid Levels and Total and Cause-specific Mortality: Prospective Evidence from 14 Cohorts in the Fatty Acids and Outcomes Research Consortium. *Circulation* 141, A43. doi:10.1161/circ.141

Tintle, N. L., Pottala, J. V., Lacey, S., Ramachandran, V., Westra, J., Rogers, A., et al. (2015). A Genome-wide Association Study of Saturated, Mono- and Polyunsaturated Red Blood Cell Fatty Acids in the Framingham Heart Offspring Study. *Prostaglandins, Leukot. Essent. Fatty Acids* 94, 65–72. doi:10.1016/j.plefa.2014.11.007

von Berg, J., van der Laan, S. W., McArdle, P. F., Malik, R., Kittner, S. J., Mitchell, B. D., et al. (2020). Alternate Approach to Stroke Phenotyping Identifies a Genetic Risk Locus for Small Vessel Stroke. *Eur. J. Hum. Genet. EJHG* 28, 963–972. doi:10.1038/s41431-020-0580-5

Wolf, J. M., Barnard, M., Xia, X., Ryder, N., Westra, J., and Tintle, N. (2020). Computationally Efficient, Exact, Covariate-Adjusted Genetic Principal Component Analysis by Leveraging Individual Marker Summary Statistics from Large Biobanks. *Pac. Symp. Biocomputing* 25, 719–730. doi:10.1142/9789811215636

Wolf, J. M., Westra, J., and Tintle, N. (2021). Using Summary Statistics to Evaluate the Genetic Architecture of Multiplicative Combinations of Initially Analyzed Phenotypes with a Flexible Choice of Covariates. bioRxiv. doi:10.1101/2021.03.08.433979

Wu, P., Wang, B., Lubitz, S. A., Benjamin, E. J., Meigs, J. B., and Dupuis, J. (2021). Approximate Conditional Phenotype Analysis Based on Genome Wide Association Summary Statistics. *Scientific Rep.* 11, 2518. doi:10.1038/s41598-021-82000-1

Zhu, X., Feng, T., Tayo, B., Liang, J., Young, J., Franceschini, N., et al. (2015). Meta-analysis of Correlated Traits *via* Summary Statistics from GWASs with an Application in Hypertension. *Am. J. Hum. Genet.* 96, 21–36. doi:10.1016/j.ajhg.2014.11.011

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.