Published in final edited form as:

Stat Appl Genet Mol Biol. ; 11(5): . doi:10.1515/1544-6115.1790.

Analyzing Genetic Association Studies with an Extended Propensity Score Approach

Huaqing Zhao,

Biostatistics and Data Management Core, The Children's Hospital of Philadelphia and Westat

Timothy R. Rebbeck, and

Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine

Nandita Mitra

Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine

Abstract

Propensity scores are commonly used to address confounding in observational studies. However, they have not been previously adapted to deal with bias in genetic association studies. We propose an extension of our previous method (Zhao et al., 2009) that uses a multilevel propensity score approach and allows one to estimate the effect of a genotype under an additive model and also simultaneously adjusts for confounders such as genetic ancestry and patient and disease characteristics. Using simulation studies, we demonstrate that this extended genetic propensity score (eGPS) can adequately adjust and consistently correct for bias due to confounding in a variety of circumstances. Under all simulation scenarios, the eGPS method yields estimates with bias close to 0 (mean=0.018, standard error=0.01). Our method also preserves statistical properties such as coverage probability, Type I error, and power. We illustrate this approach in a population-based genetic association study of testicular germ cell tumors and KITLG and SPRY4 susceptibility genes. We conclude that our method provides a novel and broadly applicable analytic strategy for obtaining less biased and more valid estimates of genetic associations.

Keywords

population-based genetic association; propensity scores; population stratification; confounding; genetic and non-genetic covariates; susceptibility genes

1 Introduction

It is important in genetic association studies to recognize and adjust for genetic variability in the study population which may otherwise lead to biased associations. Furthermore, large differences in patient characteristics in the study population, if not properly controlled for, can severely bias the estimated genetic association. Genetic studies are uniquely susceptible to bias due to confounding by population stratification (PS). PS is a type of confounding that can arise when differences in allele frequencies and differences in baseline disease rates in a population of mixed racial/ethnic groups confound the association of genes with disease. Because the frequency of many genotypes varies across races, and disease outcomes also vary by race, PS may lead to either false-positive or false-negative associations (Rebbeck, 2006).

Current methods for testing and/or adjusting confounding by PS can be broadly classified into two groups: (1) methods for controlling confounding which include genomic control (GC) (Devlin and Roeder, 1999; Devlin et al., 2001) and structured association (SA) (Pritchard and Rosenberg, 1999; Pritchard et al., 2000; Hoggart et al., 2003; Pritchard and

Donnelly, 2001); and (2) methods for constructing genomic variables that can then be used for controlling confounding as part of another approach including principal component analysis (PCA) (Price et al., 2006; Patterson et al., 2006), the stratification score (StratScore) (Epstein et al., 2007; Allen and Satten, 2011), multidimensional scaling (MDS) (Li and Yu, 2008), the genomic propensity score (GPS) (Zhao et al., 2009), spectral graph theory (Spectral-GEM) (Lee et al., 2010a; Lee et al., 2010b), and the genotype-associated propensity score (GAPS) (Lin and Zeng, 2011). However, each has its own limitations which we briefly discuss below. GC evaluates whether confounding due to PS exists by calculating an inflation factor, which is defined as the median χ^2 association test statistic divided by its theoretical median under the null distribution. An inflation factor > 1 indicates the presence of population stratification. One of the main assumptions of GC is that if the study population comes from a larger population comprised of a mixture of subpopulations with different disease prevalences and disease allele frequencies, then the χ^2 association test statistic follows a non-central χ^2 distribution (Pritchard and Rosenberg, 1999). If the non-centrality parameter is truly large, then any adjustment using the estimated inflation factor will not be sufficient to prevent false positive associations and may result in a loss of statistical power (Chen et al., 2003). SA clusters individuals into discrete subpopulations based on allele frequencies and controls for population structure. This approach has two limitations: results are highly dependent on the number of clusters and realistic populations do not naturally resolve into discrete clusters. If fractional membership in more than one cluster is allowed, the calculations become computationally intractable for the large data sets currently available (Lee et al., 2010a). PCA is a linear dimensionality reduction method to identify a small number of principal components that capture most of the relevant structure from a larger number of genetic markers. However, determining the number of significant principal components is challenging and it is common to arbitrarily use 10 principal components. In addition, PCA is sensitive to outliers (Luca et al., 2008). Due to outliers, numerous dimensions of ancestry appear to model a statistically significant amount of variation in the data, but in actuality they function to separate a single observation from the bulk of the data (Lee et al., 2010a). Similar to PCA, MDS adjusts for the confounding effect of population structure using a linear combination of principal coordinates as well as group memberships. MDS is an extension of PCA; therefore MDS and PCA share some similar limitations. An additional point of concern in MDS is the lack of a unique similarity metric that quantifies the genetic relationship between two subjects. Another method, StratScore, computes a stratification score by performing partial least-squares regression of case-control status on a matrix of genotypes at markers used to correct for ancestry. The stratification score is then used to group subjects into a number of strata. The association between case or control status and the genotype of interest is then evaluated using the stratified data. The original StratScore has been extended recently to accommodate estimation and to develop the theory underlying the StratScore which has been shown to have many of the properties of a propensity score. The main drawback for the StratScore method is that it can only be applied to case-control studies, and is not applicable for cohort or case-cohort studies. Also, the power of the StratScore approach decreases dramatically as the number of ancestry markers increases, thereby restricting the number of markers that can be used (Lee et al., 2008). Our previously developed GPS method uses the propensity score to correct for bias due to PS that considers both genetic and non-genetic factors. GPS calculates a propensity score by regressing the genotype of each candidate SNP on ancestry-informative markers (AIMs) and patient characteristics, and uses the resulting propensity score to correct for bias due to PS. However, our method at that time was limited to a two-category genetic model (e.g., a dominant genetic model). Spectral-GEM is an approach that draws from ideas across PCA, MDS and spectral graph theory. Spectral-GEM utilizes a spectral embedding derived from the so-called normalized Laplacian of a graph. The main concern of the spectral-GEM approach is the non-unique choice of weight matrix, which reflects the strength of the connections between pairs of subjects. Any matrix that measures the non-negative genetic

similarity between pairs of individuals is a candidate. Spectral-GEM only demonstrates a special weight matrix that utilizes a simple kernel derived from the traditional kernel employed in PCA analysis, and further study is required to develop a data-dependent choice of the weight matrix (Lee et al., 2010a). The GAPS approach is essentially our GPS approach but provides a natural extension to GWAS data.

To overcome some of the limitations in these current methods, we now propose an extended GPS (eGPS) approach that allows one to estimate the effect of a more general and natural three-level genetic model (e.g. an additive model) that adjusts for confounders including random null markers, AIMS, patient characteristics, and other risk factors. Specifically, our method allows for three category genotypes using an extended propensity score (Imai and Van Dyk, 2004) so that estimating the average genetic effect is possible with several levels while still maintaining the advantages of Rosenbaum and Rubin's propensity score. We describe the eGPS method in detail and we study the operating characteristics of our extended propensity score approach using simulation studies to estimate the effect of a genotype under an additive genetic model and simultaneously adjust for other confounders. We apply our method to a population-based genetic association study of testicular germ cell tumors and *KITLG* and *SPRY4* susceptibility genes.

2 Methods

The *propensity score*, defined as the conditional probability of assignment to a particular group given a vector of observed covariates, provides an appealing approach to address confounding (Rosenbaum and Rubin, 1983). Because exact adjustment using the propensity score will on average remove all of the bias in group effect estimates (Rosenbaum and Rubin, 1985), propensity scores can be used as a basis for matching, stratification, regression adjustment, and data reduction. The method offers distinct advantages because it can control for numerous covariates simultaneously by matching or stratifying on a single scalar variable; this greatly simplifies model building and estimation. Rosenbaum and Rubin's original propensity score methodology was developed exclusively for situations where there are only two treatment groups. Others have extended these methods to accommodate multi-valued and continuous treatments such as Joffe and Rosenbaum (1999), Imbens (2000), Lechner (2001), Imai and van Dyk (2004) and Hirano and Imbens (2004). Based on these methods, closely following the approach by Imbens (2000), we develop our extended genetic propensity score below.

2.1 Definitions and Theorems underlying the Extended Genetic Propensity Score Approach

We use G to denote the genotype of interest (e.g. there are potentially three genotypes: AA, AG, GG) and $\mathbf{X}$ to denote a vector of genetic covariates such as random null markers and ancestry informative markers and non-genetic covariates which may include patient demographics and disease-specific variables such as tumor characteristics.

Notation—The genotype G can take on values within the set $\mathfrak{G} = \{0, 1, 2\}$, where $\mathfrak{G}$ is the set of possible groups. Let $\mathbf{X}$ be the vector of observed covariates. Let $Y_i(g)$ $i = 1, \dots, n$ be the potential response for subject i . By potential response, we mean, for example, that $Y_i(1)$ is the value of the response that would be observed if subject i had genotype 1, and $Y_i(2)$ is the value of the response that would be observed *on the same subject* if he had genotype 2 (Holland 1986). Next, we define $I_i(g)$ to be the indicator variable, for subject i , of receiving group assignment g :

$$I_i(g) = \begin{cases} 1 & \text{if } G_i = g \\ 0 & \text{if } G_i \neq g \end{cases}$$

Definition: The Extended Propensity Score—The extended propensity score $\lambda(g, x)$ is defined to be the probability of being assigned to group $g \in \mathfrak{G}$ conditional on the observed covariates:

$$\lambda(g, x) = \Pr(G=g|x) = E[I(g)|x]$$

Using definitions of $I(g)$ and $\lambda(g, x)$, we can now give the following theorem.

THEOREM 1: The group assignment indicator and the observed covariates are conditionally independent given the extended propensity score (balancing property).

That is,

$$x \perp I(g) | \lambda(g, x)$$

THEOREM 2: If the group assignment is unconfounded given the observed covariates, then it is unconfounded given the extended propensity score (unconfoundedness).

That is,

$$\text{If } Y(g) \perp I(g) | x, \text{ then } Y(g) \perp I(g) | \lambda(g, x)$$

We have now presented two important properties of the extended genetic propensity score. First, we demonstrated the balancing property of the extended propensity score. Secondly, we showed the conditional independence of the potential response and group assignment given the extended propensity score. These two theorems play a vital role in the usefulness of the extended genetic propensity score in obtaining an unbiased estimate of genetic associations as described in section 2.2 below.

2.2 The Extended Genetic Propensity Score Approach

We use G to denote the candidate genotype of interest, D to denote disease status, X to denote a vector of genetic covariates and other non-genetic covariates which may include patient demographics and disease-specific variables. For all terms, the indices i , k , and g represent individual, subpopulation, and test-locus genotype, respectively. For example, p_{gik} represents the probability of having a high-risk genotype g for individual i in subpopulation k .

We define the *extended genetic propensity score (eGPS)* to be the likelihood of an individual having a particular test-locus genotype based on that individual's covariate makeup. This can be stated explicitly as

$$eGPS_i(g_i, x_i) = P(G_i | x_i)$$

where $eGPS_i(g_i, x_i)$ is the genomic propensity score for subject i calculated from that subject's x_i which represents that individual's vector of genetic and non-genetic covariates

and where G_i is that subject's test-locus genotype. For example, for an additive disease risk allele A and reference allele a , $G = 2$ may represent the high-risk genotype AA , $G = 1$ may represent the heterozygote risk genotype Aa , and $G = 0$ may represent the homozygote wild-type genotype aa .

In the development below, we suppress an index i corresponding to the i th individual. Based on Theorem 2 in section 2.1, we can assume that given $eGPS(g, x)$, (X) and G are conditionally independent,

$$\text{i.e., } \Pr(X, G | eGPS(g, x)) = \Pr(X | eGPS(g, x)) \Pr(G | eGPS(g, x)),$$

which balances measured covariates across test-locus genotype groups.

We define a general class of models that specify the potential relation among disease, test-locus genotype, genetic and non-genetic covariates.

$$f(E(D|X)) = \eta$$

where $f(\cdot)$ is a link function, such as the logit function, that determines the relationship between the outcome variable D and predictor variable X ; $E(D|X)$ denotes the conditional mean of D given X ; and η is a function of covariates, usually a linear function. For example, we assumed that the true disease prevalence follows logit link for our simulations:

$$\eta = \text{logit}(D=1|X, G) = \alpha + \beta G + \beta_d X$$

where $\text{logit}(t) = \log(t/(1-t))$ is the logit function; α is baseline odds-ratio between (X, G) and D ; β is log odds-ratio between $G|X$ and D ; β_d is log odds-ratio between $X|G$ and D .

Then, under the assumptions of balance scoring and unconfoundedness of the propensity score, we specify our proposed eGPS model:

$$\eta = \alpha + \beta G + \gamma_1 eGPS(G=1, X) + \gamma_2 eGPS(G=2, X)$$

where α and β are log odds ratios and γ_1 and γ_2 are nuisance parameters.

We fit the disease prevalence model assuming a logit link to estimate the effect of risk genotype(s) β :

$$\text{Logit}(D|G, eGPS(G, X)) = \alpha + \beta G + \gamma_1 eGPS(G=1, X) + \gamma_2 eGPS(G=2, X).$$

3 Simulation Studies

We conducted Monte Carlo simulation studies to evaluate the performance of the proposed eGPS method compared to PCA and MDS and to compare it with a model with no adjustment for non-genetic and genetic covariates. In summary, data for a population consisting of 2 subpopulations were generated. The allele frequency p_k for the k th subpopulation was randomly sampled from $p_k \sim \text{Beta}[(1-F_{st})p/F_{st}, (1-F_{st})(1-p)/F_{st}]$, where $k=1$ or 2 for subpopulation, $p=0.1, 0.3$, or 0.5 for minor allele frequency (MAF), and $F_{st}=0.01, 0.05$, or 0.1 for Wright's coefficient of inbreeding. F_{st} is a measure of the genetic distance between two populations that describes the proportion of overall genetic variation

that is due to differences between populations. Samples of 500 cases and 500 controls were drawn from a population consisting of two subpopulations. The prior probability of sampling individuals from each subpopulation was 0.5. The prevalence of the disease was varied from 0.25 to 0.75 in subpopulations 1 and 2. Under no PS, the prevalence of the disease was the same in each subpopulation. Thus, the expected numbers of cases and controls from each subpopulation were (250, 250) and (250, 250), respectively. Under moderate PS, the prevalence of the disease was 1.5-fold higher in subpopulation 1; thus, the expected numbers of cases and controls for each subpopulation were (300, 200) and (200, 300), respectively. Finally, under severe PS, the prevalence of the disease was 3-fold higher in subpopulation 1; thus, the expected number of cases and controls from each subpopulation were (375, 125) and (125, 375), respectively. Finally, the log odds ratio between G and D was chosen to be $\log(1)$ and $\log(1.5)$, i.e., $OR=1$ for no genetic association and 1.5 indicating genetic association.

For the estimation of parameters, we used a two-step process. First, to estimate the extended genomic propensity $eGPS(X)$, we used a multinomial logistic regression model to obtain the predicted probability of having the number of risk alleles based on X . The dependent variable was the test-locus genotype and the independent variables were the covariates. Second, to estimate the effect of the genotype on disease, we used a second multinomial logistic regression model. In this case, the dependent variable was the disease status, and the independent variables were the test-locus genotype variable along with the estimated $eGPS$.

For each scenario, we generated 1,000 simulation replicates and evaluated the odds ratio (OR), bias, standard error (SE) of the estimated log OR, the empirical coverage probability (CP), Type I error, and power. We compared the performance of the $eGPS$ method to other methods with respect to bias and coverage probability. Finally, we evaluated Type I error and power for our $eGPS$ approach.

3.1 Simulation Results

Table 1 shows the results comparing the no adjustment, PCA, MDS, GPS and $eGPS$ methods with $OR=1$ in terms of bias, coverage probability, and Type I error. As expected, under no/mild PS, all four adjusted methods performed well in terms of bias, CP and Type I error. Even under no/mild PS, all adjusted methods consistently outperform the no adjustment method in terms of bias, CP and Type I error.

Under moderate PS, the no adjustment method yielded overestimation (bias ranging from 0.39 to 0.468) for all combinations of F_{st} , $p1$ and $p2$. Without adjustment, the coverage probability was almost zero with a range from 0 to 0.004 under moderate PS. Our proposed $eGPS$ method consistently outperforms the PCA and GPS methods in terms of bias, CP and Type I error under moderate PS. MDS sometimes outperformed $eGPS$ in terms of bias and CP under moderate PS, though inconsistently. However, our $eGPS$ method consistently outperformed the MDS method in terms of Type I error. As expected, the basic no adjustment method yielded an inflated Type I error rate under moderate PS.

Similarly, under severe PS, no adjustment performed poorly, yielded high bias, zero CP and 100% Type I error rate. The $eGPS$ method consistently outperformed the PCA and GPS methods in terms of bias, CP and Type I error under severe PS. MDS outperformed $eGPS$ in terms of bias most times under severe PS. However, our $eGPS$ method mostly outperformed the MDS method in terms of CP and Type I error under severe PS. Note that for the combination of $F_{st}=0.01$, $p1=0.1$ and $p2=0.3$, all four adjustment methods performed poorly under severe PS. We further estimated the inflation factors for all combinations of F_{st} , $p1$ and $p2$ under severe PS. The estimated inflation factor is 30 for $F_{st}=0.01$, $p1=0.1$ and

$p3=0.3$, but more than 60 (range from 64 to 200) for all other scenarios under severe PS in Table 1.

Table 2 shows the results of comparing the no adjustment, PCA, MDS, GPS and eGPS methods with $OR=1.5$ in terms of bias, coverage probability, and power. As expected, under no/mild PS, all methods performed well in terms of low bias, high CP and power. Under the moderate PS, all four adjusted methods performed reasonably well in terms of bias, although the eGPS method yielded the lowest bias for most F_{st} and MAF combinations. The proposed eGPS method consistently outperforms the PCA, MDS and GPS methods in terms of CP regardless of F_{st} and MAF. Under moderate PS, all methods yielded high power. Under severe PS, the proposed eGPS consistently outperformed both PCA and GPS methods in terms of bias and CP. eGPS sometimes outperformed MDS in terms of bias and CP under severe PS, though inconsistently. Finally, all methods yielded high power under severe PS. Note that for the combination of $F_{st}=0.01$, $p1=0.1$ and $p3=0.3$ under severe PS, all four adjustment methods performed poorly in terms of bias and CP. We further estimated the inflation factors for all combinations of F_{st} , $p1$ and $p3$ under severe PS. The estimated inflation factor is 8.6 for $F_{st}=0.01$, $p1=0.1$ and $p3=0.3$, but 29 to 150 for all other scenarios under severe PS in Table 2. For continuity, we still keep this combination under the severe PS situation. This indicates that all adjustment methods may not appropriately apply a greater correction to markers with relatively less inflated population stratification.

Figures 1 and 2 illustrate results of comparing the no adjustment, PCA, MDS, GPS and eGPS methods with $OR=1.5$ in terms of bias and CP by MAF for $F_{st} = 0.01, 0.05$, or 0.1 under moderate PS. Figure 1 shows that all four adjusted methods perform reasonably well in terms of bias, although eGPS performs consistently better in terms of removing bias due to PS. This is also true for CP as shown in Figure 2.

4 Application to a Genetic Association Study of Testicular Cancer

Testicular germ cell tumors (TGCT) are the most common cancers in young men in the US, with the incidence increasing in white men. It is postulated that there is a substantial genetic contribution to TGCT. In 2009, a genome-wide association study (GWAS) was conducted among 292 cases with TGCT and 919 controls. We applied our methods to the top six statistically significant SNPs from this GWAS (Kanetsky et al., 2009). All 292 cases and 919 controls were self-reported whites. However, 15 cases had Asian or African ancestry as determined by multidimensional scaling using all genome-wide markers. We used the extended genetic propensity score to analyze the top 6 GWAS SNPs (SPRED2: rs17031166, rs1549383; SPRY4: rs4324715, rs6897876; KITLG: rs3782179, rs4474514) assuming an additive genetic model. We selected 440 null markers from across the genome. These 440 null markers were selected so that they were not associated with disease and were not in linkage disequilibrium with each other.

To apply our approach to the TGCT dataset, we used a two-step process to analyze the six selected SNP disease markers. First, to estimate the two extended genetic propensity scores (eGPS), we used a multinomial logistic regression model to obtain the predicted probability of having the number of risk alleles for each significant SNP based on the 440 null markers. Next, to estimate the effect of each SNP on disease, we used a second logistic regression model. Here, the dependent variable was disease status, and the independent variables were the top SNP along with the two estimated extended genetic propensity scores. We repeated the process six times to analyze the top 6 GWAS SNPs.

Table 3 shows the summary of associations between the selected SNP markers and testicular germ cell tumor (TGCT). TGCT risk was increased 54% per copy of the major (G) allele at rs17031166 with no adjustment. After excluding 15 non-white TGCT cases as determined

by MDS, TGCT risk was increased to 60% per copy of the major (G) allele at rs17031166. However, the TCGT risk was only increased to 34% per copy of the major allele (G) at rs17031166 after adjustment with the eGPS. For all six SNP markers, the increased risk of TGCT per risk allele was inflated by about 10 to 20% per allele without adjustment, and 20 to 60% per allele with MDS adjustment in comparisons to eGPS adjustment. This suggests that after adjusting for the eGPS, TGCT risk estimates are attenuated towards the null for all six SNP markers. This phenomenon was also observed in our simulation studies; for example, when the true OR was 1.5 and there was moderate population stratification, the eGPS approach slightly underestimated the OR for the genetic effect.

We further analyzed the data by excluding 15 non-white TGCT cases as determined by MDS with eGPS adjustment on the sample of N=1196. This sample was selected by using the MDS approach based on all available genome-wide markers. With this more homogenous population, the eGPS method consistently reduced residual confounding when compared with no adjustment. It suggests that eGPS adjustment can be of benefit in reducing potential false positives even in the presence of mild population stratification.

5 Conclusion

Propensity scores are commonly used to address confounding in observational studies. The original work of Rosenbaum and Rubin (1983) considered applications with binary treatments. We provide a unique application to an extension of these methods (Imbens, 2000) to genetic association studies. Specifically, we develop and demonstrate the usefulness of the extended genetic propensity score in the context of 3-level genotypes in order to address population stratification bias. Currently, commonly applied methods control bias due to confounding by PS using genetic markers alone. Our eGPS method, which utilizes the estimated extended propensity score, allows us to represent each subject's genetic and non-genetic background with a single variable and captures much of the variation due to genetic as well as non-genetic factors. As illustrated in our simulation studies, the eGPS method can adjust adequately and consistently for confounding by PS when there are moderate and severe ancestral differences between cases and controls. Our simulations also demonstrate that eGPS preserves statistical properties such as coverage probability, Type I error, and power. Our encouraging results from both the simulation study and our real data example suggest that the eGPS method provides a novel and robust tool for obtaining less-biased estimates of genetic associations.

We have developed and applied these methods in the context of case-control studies. However, as pointed out by Joffe and Rosenbaum (1999) and recently re-emphasized by Allen et al. (2010), the propensity score is traditionally used to adjust for confounding in a cohort or case-cohort study. Propensity scores are also readily applicable to nested case-control studies (Rosenbaum 2010, personal communication). However, the valid use of propensity scores in typical case-control studies has received little attention and is less straightforward than their use in cohort studies (Mansson et al., 2007). Mansson et al. present theory and simulations on the estimation and use of propensity score in case-control and case-cohort studies to address two potential issues: 1) artificial effect modification by the estimated propensity score; and 2) residual confounding due to bias in the estimation of the true propensity score. They demonstrated that little to no effect modification was induced when using propensity score adjustment to deal with confounding in case-control studies. The magnitude of residual confounding due to bias in the estimation of the true propensity score appeared modest in these simulations. In our simulation studies, it is also clearly demonstrated that potential residual confounding has been adequately removed by use of the propensity score. Recently, Lin and Zeng (2011) have provided theoretical justification for using the propensity score in a logistic regression model.

Our simulation studies might have limitations based on the current range of parameter selections and combinations. For example, all adjustment methods performed poorly, yielded high bias, low CP and high Type I error rate with $F_{st}=0.01$, $p1=.1$ and $p2=.3$ under severe population stratification. However, it is unlikely that such a high inflation factor would occur in a carefully designed study. Our choice of simulation parameters reflect realistic ranges that we have experienced in practice in our own research.

We present our method in the context of a binary disease outcome. It is straightforward to extend the proposed approach to study other traits, such as quantitative traits, by using a generalized linear model approach (McCullagh and Nelder, 1989). In addition, we have presented our approach in the context of a candidate gene association study; we are currently extending our eGPS method for application to genome-wide association studies by combining principal component analysis and the eGPS.

An additional advantage to our method is that it is computationally less burdensome and can be easily implemented by using standard software such as R and SAS, unlike other PS methods that require specialized software. In summary, our eGPS method allows one to easily adjust for genetic and non-genetic confounders, without specialized software, to adjust for population stratification in population-based genetic association studies.

Acknowledgments

The authors thank Drs. Peter A. Kanetsky and Katherine L. Nathanson for providing the genetic data from their testicular germ cell cancer study, and Saran Vardhanabhuti for creating data files for the candidate and null markers. TRR was funded in part by grants P50-CA105641 and R01-CA08574 from NIH. NM was funded in part by a grant with the Pennsylvania Department of Health. The Department specifically disclaims responsibility for any analysis, interpretations or conclusions.

6 Appendix

Using definitions of $I(g)$ $\lambda(g, x)$, we can now prove theorems 1 and 2.

THEOREM 1

The group assignment indicator and the observed covariates are conditionally independent given the extended propensity score (balancing property).

That is,

$$x \perp I(g) | \lambda(g, x)$$

First note that

$$\begin{aligned} \Pr \{I(g)=1|x, \lambda(g, x)\} \\ &= E[I(g)|x, \lambda(g, x)] \\ &= E[I(g)|x] \\ &= \lambda(g, x) \end{aligned}$$

Now,

$$\begin{aligned}
& \Pr \{I(g)=1|\lambda(g, x)\} \\
&= E[I(g)|\lambda(g, x)] \\
&= E \{E[I(g)|x, \lambda(g, x)]|\lambda(g, x)\} \\
&= E[\lambda(g, x)|\lambda(g, x)] \\
&= \lambda(g, x)
\end{aligned}$$

Hence,

$$\begin{aligned}
\Pr \{I(g)=1|x, \lambda(g, x)\} &= \Pr \{I(g)=1|\lambda(g, x)\} \\
&\Rightarrow x \perp I(g)|\lambda(g, x).
\end{aligned}$$

THEOREM 2

If the group assignment is unconfounded given the observed covariates, then it is unconfounded given the extended propensity score (unconfoundedness).

That is,

$$\text{If } Y(g) \perp I(g)|x, \text{ then } Y(g) \perp I(g)|\lambda(g, x)$$

First note that

$$\begin{aligned}
& \Pr \{I(g)=1|Y(g), \lambda(g, x)\} \\
&= E[I(g)|Y(g), \lambda(g, x)] \\
&= E \{E[I(g)|Y(g), x, \lambda(g, x)]|Y(g), \lambda(g, x)\} \\
&= E[\lambda(g, x)|Y(g), \lambda(g, x)] \text{ since } Y(g) \perp I(g)|x \\
&= \lambda(g, x)
\end{aligned}$$

Now,

$$\begin{aligned}
& \Pr \{I(g)=1|\lambda(g, x)\} \\
&= E[I(g)|\lambda(g, x)] \\
&= E \{E[I(g)|x, \lambda(g, x)]|\lambda(g, x)\} \\
&= E[\lambda(g, x)|\lambda(g, x)] \\
&= \lambda(g, x)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\Pr \{I(g)=1|Y(g), \lambda(g, x)\} &= \Pr \{I(g)=1|\lambda(g, x)\} \\
&\Rightarrow Y(g) \perp I(g)|\lambda(g, x).
\end{aligned}$$

References

- Allen AS, Epstein MP, Satten GA. Score-based adjustment for confounding by population stratification in genetic association studies. *Genetic Epidemiology*. 2010; 34(5):383–385. [PubMed: 20127852]

- Allen AS, Satten GA. Control for confounding in case-control studies using the stratification score, a retrospective balancing score. *American Journal of Epidemiology*. 2011; 173(7):752–760. [PubMed: 21402731]
- Chen HS, Zhu X, Zhao H, Zhang S. Qualitative semi-parametric test for genetic associations in case-control designs under structured populations. *Annals of Human Genetics*. 2003; 67:250–264. [PubMed: 12914577]
- Devlin B, Roeder K. Genomic control for association studies. *Biometrics*. 1999; 55:997–1004. [PubMed: 11315092]
- Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. *Theoretical Population Biology*. 2001; 60:155–166. [PubMed: 11855950]
- Epstein MP, Allen AS, Satten GA. A Simple and Improved Correction for Population Stratification in Case-Control Studies. *The American Journal of Human Genetics*. 2007; 80:921–930.
- Hirano, K.; Imbens, GW. The Propensity Score with Continuous Treatments. In: Gelman, A.; Meng, X-L., editors. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* Chichester. England: John Wiley & Sons; 2004.
- Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG, McKeigue PM. Control of confounding of genetic associations in stratified populations. *The American Journal of Human Genetics*. 2003; 72:1492–1504.
- Holland PW. Statistics and causal inference. *Journal of the American Statistical Association*. 1986; 81:945–970.
- Imai K, Van Dyk DA. Cause inference with general group regimes: generalizing the propensity score. *Journal of the American Statistical Association*. 2004; 99:854–866.
- Imbens GW. The Role of the Propensity Score in Estimating Dose-Response Functions. *Biometrika*. 2000; 87:706–710.
- Joffe MM, Rosenbaum PR. Invited commentary: propensity scores. *The American Journal of Epidemiology*. 1999; 150:327–333.
- Kanetsky PA, Mitra N, Vardhanabhuti S, Li M, Vaughn DJ, Letrero R, Ciosek SL, Doody DR, Smith LM, Weaver J, Albano A, Chen C, Starr JR, Rader DJ, Godwin AK, Reilly MP, Hakonarson H, Schwartz SM, Nathanson KL. Common variation in KITLG and at 5q31.3 predisposes to testicular germ cell cancer. *Nature Genetics*. 2009; 41(7):811–815. [PubMed: 19483682]
- Lechner, M. Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In: Lechner, M.; Pfeiffer, F., editors. *Econometric Evaluation of Labour Market Policies*. New York: Physica-Verlag Heidelberg; 2001.
- Lee AB, Luca D, Klei L, Devlin B, Roeder K. Discovering genetic ancestry using spectral graph theory. *Genetic Epidemiology*. 2010a; 34:51–59. [PubMed: 19455578]
- Lee AB, Luca D, Roeder K. A spectral graph approach to discovering genetic ancestry. *The Annals of Applied Statistics*. 2010b; 4(1):179–202. [PubMed: 20689656]
- Lee S, Sullivan PF, Zou F, Wright FA. Comment on simple and improved correction for population stratification. *The American Journal of Human Genetics*. 2008; 82:524–531.
- Li QZ, Yu K. Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. *Genetic Epidemiology*. 2008; 32:215–226. [PubMed: 18161052]
- Lin DY, Zeng D. Correcting for population stratification in genomewide association studies. *Journal of the American Statistical Association*. 2011; 106(495):997–1008. [PubMed: 22467997]
- Luca D, Ringquist S, Klei L, Lee A, Gieger C, Wichmann HE, Schreiber S, Krawczak M, Lu Y, Styche A, Devlin B, Roeder K, Trucco M. On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. *The American Journal of Human Genetics*. 2008; 82:453–463.
- Mansson R, Sun W, Joffe MM, Hennessy S. On the estimation and use of propensity scores in case-control and case cohort studies. *American Journal of Epidemiology*. 2007; 165:1–8. [PubMed: 17074966]
- McCullagh, P.; Nelder, JA. *Generalized Linear Model*. Second Edition. New York: Chapman & Hall/CRC Press; 1989.

- Patterson NJ, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genetics*. 2006; 2(12):e190. [PubMed: 17194218]
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*. 2006; 38:904–909. [PubMed: 16862161]
- Pritchard JK, Donnelly P. Case-control studies of association in structured or admixed populations. *Theoretical Population Biology*. 2001; 60:227–37. [PubMed: 11855957]
- Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. *The American Journal of Human Genetics*. 1999; 65:220–228.
- Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. *The American Journal of Human Genetics*. 2000; 67:170–181.
- Rebbeck TR. Inherited genetic markers and cancer outcomes: personalized medicine in the postgenome era. *Journal of Clinical Oncology*. 2006; 24(13):1972–1974. [PubMed: 16648495]
- Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983; 70:41–55.
- Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*. 1985; 39:33–38.
- Zhao H, Rebbeck TR, Mitra N. A propensity score approach to correction for bias due to population stratification using genetic and non-genetic factors. *Genetic Epidemiology*. 2009; 33:679–690. [PubMed: 19353632]

Figure 1.

Bias of estimated odds ratio under moderate population stratification. Dash-dot line is no adjustment. Short dash line is PCA. Dash line is MDS. Long dash line is GPS. Solid line is eGPS.

Figure 2. Coverage probability under moderate population stratification. Dash-dot line is no adjustment. Short dash line is PCA. Dash line is MDS. Long dash line is GPS. Solid line is eGPS.

Table 1

Bias of estimated odds ratio (OR), the coverage probability (CP), and the empirical Type I error of simulations for comparing no adjustment, PCA, MDS, GPS and eGPS methods with true OR=1.

F _{st}	p1	p2	Bias					CP					Type I Error				
			No Adjustment	PCA	MDS	GPS	eGPS	No Adjustment	PCA	MDS	GPS	eGPS	No Adjustment	PCA	MDS	GPS	eGPS
No/Mild Population Stratification																	
0.01	0.1	0.3	0.185	-0.015	-0.018	0.012	-0.015	0.456	0.949	0.951	0.965	0.964	0.544	0.051	0.049	0.035	0.036
0.01	0.1	0.5	0.127	0.020	0.021	0.020	0.024	0.685	0.940	0.943	0.949	0.967	0.315	0.060	0.057	0.051	0.033
0.05	0.1	0.3	0.162	-0.056	-0.057	-0.026	-0.048	0.537	0.913	0.903	0.938	0.926	0.463	0.087	0.097	0.062	0.074
0.05	0.1	0.5	0.125	-0.001	0.005	-0.011	-0.002	0.701	0.946	0.946	0.946	0.964	0.299	0.054	0.054	0.054	0.036
0.1	0.1	0.3	0.117	0.006	0.002	0.002	0.002	0.746	0.950	0.942	0.951	0.963	0.254	0.050	0.058	0.049	0.037
0.1	0.1	0.5	0.108	-0.011	-0.026	-0.004	-0.011	0.788	0.952	0.946	0.960	0.969	0.212	0.048	0.054	0.040	0.031
Moderate Population Stratification																	
0.01	0.1	0.3	0.390	0.051	0.027	0.082	0.044	0.003	0.915	0.942	0.884	0.937	0.997	0.085	0.058	0.116	0.063
0.01	0.1	0.5	0.460	0.042	0.021	0.079	0.035	0.000	0.910	0.953	0.941	0.962	1.000	0.059	0.047	0.090	0.038
0.05	0.1	0.3	0.393	-0.026	-0.059	0.012	-0.024	0.004	0.953	0.918	0.947	0.953	0.996	0.053	0.082	0.047	0.047
0.05	0.1	0.5	0.468	0.022	0.005	0.041	0.015	0.000	0.941	0.953	0.944	0.958	1.000	0.056	0.047	0.059	0.042
0.1	0.1	0.3	0.428	0.019	0.001	0.036	0.009	0.000	0.960	0.958	0.953	0.974	1.000	0.047	0.042	0.040	0.026
0.1	0.1	0.5	0.434	0.017	-0.026	0.040	0.006	0.000	0.944	0.945	0.937	0.958	1.000	0.063	0.055	0.056	0.042
Severe Population Stratification																	
0.01	0.1	0.3	0.714	0.169	0.116	0.208	0.155	0.000	0.508	0.833	0.699	0.728	1.000	0.301	0.167	0.492	0.272
0.01	0.1	0.5	1.010	0.085	0.028	0.177	0.068	0.000	0.897	0.931	0.681	0.930	1.000	0.103	0.069	0.319	0.070
0.05	0.1	0.3	0.759	0.027	-0.056	0.083	0.024	0.000	0.899	0.927	0.931	0.950	1.000	0.069	0.073	0.101	0.050
0.05	0.1	0.5	1.039	0.056	0.000	0.130	0.037	0.000	0.839	0.947	0.930	0.948	1.000	0.070	0.053	0.161	0.052
0.1	0.1	0.3	0.935	0.044	0.003	0.099	0.025	0.000	0.872	0.941	0.923	0.961	1.000	0.077	0.059	0.128	0.039
0.1	0.1	0.5	0.973	0.062	-0.025	0.124	0.040	0.000	0.834	0.938	0.921	0.951	1.000	0.079	0.062	0.166	0.049

OR=odds ratio between disease (D) and gene (G); the bias is the average over 1,000 replications of the bias of estimated OR; bias=log(estimated OR)-log(true OR); the empirical coverage probability is the proportion of the confidence interval of the estimated OR including the true OR among 1,000 replications. In all models, true OR=1.

Table 2

Bias of estimated odds ratio (OR), the coverage probability (CP), and the empirical Type I error of simulations for comparing no adjustment, PCA, MDS, GPS and eGPS methods with true OR=1.5.

F _{st}	p1	p2	Bias					CP					Power				
			No Adjustment	PCA	MDS	GPS	eGPS	No Adjustment	PCA	MDS	GPS	eGPS	No Adjustment	PCA	MDS	GPS	eGPS
No/Mild Population Stratification																	
0.01	0.1	0.3	0.143	-0.003	-0.007	-0.013	-0.050	0.656	0.947	0.952	0.959	0.935	1.000	0.981	0.984	0.992	0.973
0.01	0.1	0.5	-0.036	0.043	0.027	-0.055	-0.048	0.971	0.937	0.939	0.918	0.938	0.999	0.956	0.967	0.919	0.913
0.05	0.1	0.3	0.067	0.041	0.040	-0.010	-0.031	0.906	0.936	0.933	0.962	0.958	1.000	0.988	0.988	0.988	0.972
0.05	0.1	0.5	-0.045	0.012	-0.001	-0.066	-0.066	0.953	0.948	0.952	0.908	0.918	1.000	0.943	0.955	0.911	0.904
0.1	0.1	0.3	0.220	0.005	-0.003	-0.018	-0.057	0.314	0.953	0.952	0.954	0.917	1.000	0.964	0.963	0.970	0.920
0.1	0.1	0.5	0.006	0.010	0.012	-0.069	-0.061	0.986	0.952	0.952	0.917	0.947	1.000	0.956	0.952	0.906	0.908
Moderate Population Stratification																	
0.01	0.1	0.3	0.309	0.105	0.100	0.098	0.052	0.076	0.830	0.841	0.850	0.926	1.000	1.000	1.000	1.000	0.999
0.01	0.1	0.5	0.286	0.059	0.045	0.038	-0.011	0.030	0.908	0.927	0.942	0.958	1.000	0.983	0.972	0.989	0.957
0.05	0.1	0.3	0.295	0.076	0.047	0.058	0.013	0.047	0.904	0.940	0.944	0.969	1.000	0.996	0.993	1.000	0.994
0.05	0.1	0.5	0.265	0.027	0.010	0.019	-0.034	0.074	0.947	0.954	0.964	0.961	1.000	0.968	0.955	0.978	0.939
0.1	0.1	0.3	0.432	0.069	0.042	0.062	0.022	0.001	0.899	0.927	0.917	0.953	1.000	0.994	0.984	0.995	0.982
0.1	0.1	0.5	0.329	0.044	0.054	-0.002	-0.021	0.002	0.935	0.926	0.960	0.960	1.000	0.970	0.982	0.958	0.945
Severe Population Stratification																	
0.01	0.1	0.3	0.570	0.290	0.285	0.273	0.217	0.000	0.235	0.248	0.239	0.453	1.000	1.000	1.000	1.000	1.000
0.01	0.1	0.5	0.825	0.113	0.047	0.181	0.039	0.000	0.848	0.937	0.675	0.954	1.000	0.984	0.940	0.998	0.961
0.05	0.1	0.3	0.662	0.140	0.059	0.159	0.075	0.000	0.778	0.919	0.711	0.915	1.000	0.998	0.970	1.000	0.987
0.05	0.1	0.5	0.781	0.073	0.005	0.151	0.011	0.000	0.907	0.941	0.752	0.960	1.000	0.960	0.897	0.998	0.937
0.1	0.1	0.3	0.768	0.187	0.132	0.177	0.137	0.000	0.645	0.778	0.658	0.778	1.000	1.000	0.997	0.999	0.998
0.1	0.1	0.5	0.872	0.092	0.026	0.091	0.027	0.000	0.884	0.943	0.879	0.942	1.000	0.975	0.907	0.977	0.914

OR=odds ratio between disease (D) and gene (G); the bias is the average over 1,000 replications of the bias of estimated OR; bias=log(estimated OR)-log(true OR); the empirical coverage probability is the proportion of the confidence interval of the estimated OR including the true OR among 1,000 replications. In all models, true OR=1.5.

\$watermark-text

\$watermark-text

\$watermark-text

Table 3

Risk allele frequency and per allele odds ratio (95% confidence interval) of the association between TGCT and six selected SNP disease markers.

Gene	Marker	Nonrisk/risk Allele	RAF		Per Allele OR (95% CI)		
			Controls	Cases	No Adjustment (N=1211)	MDS Adjustment (N=1196)	eGPS (N=1211)
SPRED2	rs17031166	G/C	0.30	0.39	1.54(1.27-1.88)	1.60(1.31-1.96)	1.34(1.03-1.73)
	rs1549383	A/G	0.30	0.41	1.59(1.31-1.93)	1.65(1.35-2.02)	1.44(1.11-1.87)
SPRY4	rs43245715	T/C	0.49	0.59	1.58(1.30-1.93)	1.59(1.30-1.93)	1.39(1.07-1.79)
	rs6897876	C/T	0.54	0.66	1.62(1.33-1.95)	1.59(1.31-1.94)	1.41(1.10-1.82)
KITLG	rs3782179	A/G	0.80	0.89	2.12(1.59-2.84)	2.36(1.73-3.21)	1.91(1.28-2.87)
	rs4474514	A/G	0.80	0.89	2.20(1.64-2.95)	2.45(1.79-3.35)	1.91(1.27-2.87)

TGCT, testicular germ cell tumor; Marker, dbSNP rs number; RAF, risk allele frequency; OR, odds ratio; CI, confidence interval; MDS, multidimensional scaling; eGPS, extended genomic propensity score.