Twin Research and Human Genetics (2022), **25**, 24–39
doi:10.1017/thg.2022.2

CAMBRIDGE UNIVERSITY PRESS

Article

Genetically Adjusted Propensity Score Matching: A Comparison to Discordant MZ Twin Models

Ian A. Silver[1,2], Hexuan Liu[3] and Joseph L. Nedelec[3]

[1]Law and Justice Department, Rowan University, Glassboro, NJ, USA, [2]Corrections Institute, University of Cincinnati, Cincinnati, OH, USA and [3]School of Criminal Justice, University of Cincinnati, Cincinnati, OH, USA

Abstract

Discordant monozygotic (MZ) twin methodologies are considered one of the foremost statistical approaches for estimating the influence of environmental factors on phenotypic variance. Limitations associated with the discordant MZ twin approach generates an inability to estimate particular relationships and adjust estimates for the confounding influence of gene-nonshared environment interactions. Recent advancements in molecular genetics, however, can provide the opportunity to address these limitations. The current study reviews an alternative technique, *genetically adjusted propensity scores* (GAPS) matching, that integrates observed genetic and environmental information to adjust for the confounding of these factors in nonkin individuals. Simulations and a real data example were used to compare the GAPS matching approach to the discordant MZ twin method. Although the results of the simulated comparisons demonstrated that the discordant MZ twin approach remains the more robust statistical technique to adjust for shared environmental and genetic factors, GAPS matching — under certain conditions — could represent a viable alternative when MZ twin samples are unavailable. Overall, the findings suggest that GAPS matching can potentially provide an alternative to the discordant MZ twin approach when limited variation exists between identical twin pairs. Moreover, the ability to adjust for gene-nonshared environment interactions represents a potential advancement associated with the GAPS approach. The limitations of the approach, as well as polygenic risk scores, are also discussed.

Keywords: Propensity score matching; polygenic risk scores; discordant MZ twin designs; genetically adjusted propensity scores

(Received 3 February 2022; accepted 3 February 2022; First Published online 4 May 2022)

Discordant monozygotic (MZ) twin methods can produce estimates unconfounded by genetic and shared environmental factors that is unrivaled by many other statistical methodologies (Oskarsson et al., 2017; Ross et al., 2020; Tiu et al., 2004; Vitaro et al., 2009). Since MZ twins share 100% of their genetic material and 100% of their shared environment, it can be assumed that any phenotypic differences between MZ twins correspond to divergent environmental factors (i.e., the nonshared environment; Knopik et al., 2016). As such, relative to standard social science models, the confidence in the estimates is substantively increased through the use of discordant MZ twin methods because it can be assumed that genetic and shared environmental factors are constant across MZ twins (Vitaro et al., 2009). Unobserved heterogeneity in the genome, phenotypic characteristics and shared environmental factors are adjusted for when using discordant MZ twin methods due to the co-twin establishing a robust counterfactual condition with genetic and shared environmental experiences identical to the target twin (Knopik et al., 2016).

Discordant MZ methods permit evaluation of the associations between nonshared environmental factors and phenotypic variation without the collection of molecular genetic information (e.g., single-nucleotide polymorphisms [SNPs]), information pertaining to shared phenotypes or information on the shared environment (Knopik et al., 2016). Moreover, the estimates produced by discordant MZ analyses can be adjusted for observed heterogeneity in discordant nonshared environmental factors (i.e., factors that differ between the twins; Knopik et al., 2016). Given the capacity to estimate effects with limited biases, discordant MZ twin methods are commonly used in the behavioral and social sciences and are considered robust analytical tools for evaluating causal associations at the individual level (e.g., Asbury et al., 2003; Motz et al., 2019; Oskarsson et al., 2017; Silberg et al., 2016; Thornton et al., 2017). Nevertheless, large cohorts of molecular genetic data, such as the UK Biobank, and the advancement of polygenic risk scores (PRSs) provide the ability to reduce observed genetic risk information and environmental information to a single score, potentially presenting an opportunity to match the estimates produced by discordant MZ twin methodologies using singletons.

Given the availability of molecular genetic data and advanced techniques, an alternative procedure intended to produce results similar to discordant MZ twin methods is discussed in the current study. This alternative procedure — termed *genetically adjusted propensity scores* (GAPS) — integrates polygenic scores into the calculation of a propensity score to adjust for the confounding effects of genetic and environmental factors on an association of interest through matching. The GAPS matching methodology and the results from various simulated analyses are presented. These results

Author for correspondence: Ian A. Silver, Email: silveria@rowan.edu

Cite this article: Silver IA, Liu H, and Nedelec JL. (2022) Genetically Adjusted Propensity Score Matching: A Comparison to Discordant MZ Twin Models. *Twin Research and Human Genetics* **25**: 24–39, https://doi.org/10.1017/thg.2022.2

are intended to provide an indication of the conditions necessary for GAPS matching to estimate associations within the confidence intervals of discordant MZ twin methods. Furthermore, a thorough discussion of the benefits and challenges associated with GAPS matching is provided to emphasize the validity of the methodology. GAPS matching integrates both propensity scores (created from observed environmental measures) and polygenic scores (created from observed genetic SNPs) to adjust for the confounding influence of genetic and environmental factors on associations of interest.[1]

Propensity Score Matching

Briefly, a propensity score is defined as the probability of being exposed to a treatment given the phenotype of an individual and their environmental circumstances (Guo & Fraser, 2015).[2] A propensity score can be estimated using a variety of generalized linear models, conditional upon the level of measurement of the treatment, with a binary logistic regression model being the most common (Guo & Fraser, 2015). As displayed in Eq. 1, a propensity score ($\boldsymbol{P}$) is the exponential value of a case's weighted scores ($\boldsymbol{\beta}$) on the matrix of independent variables ($\boldsymbol{X_i}$), divided by 1 plus the exponential value of a case's weighted scores ($\boldsymbol{\beta}$) on the matrix of independent variables ($\boldsymbol{X_i}$). The weights ($\boldsymbol{\beta}$) are derived from regressing the logged odds of the treatment ($\boldsymbol{v_i}$; 0 = not exposed to treatment; 1 = exposed to treatment) on the matrix of independent variables ($\boldsymbol{X_i}$):

$$\log\left(\frac{\boldsymbol{v}_i}{1-\boldsymbol{v}_i}\right) = \boldsymbol{\beta X_i} + \varepsilon \tag{1}$$

$$\boldsymbol{P} = \frac{\exp(\boldsymbol{\beta X}_i)}{1+\exp(\boldsymbol{\beta X}_i)}$$

Propensity scores can be employed in various ways to reduce the bias in the estimated effects of a treatment on an outcome of interest. A frequently used method is propensity score matching, which is a technique designed to emulate an experimental condition by matching treatment cases to control cases with a similar probability of being exposed to the treatment (i.e., propensity score; Guo & Fraser, 2015). Treatment cases and control cases can be matched using various techniques, of which the most easily understood are exact matching (Eq. 2) and caliper matching (Eq. 3; ϵ represents the caliper):

$$\boldsymbol{P}_{(t=1)} = \boldsymbol{P}_{(t=0)} \tag{2}$$

$$\left\|\boldsymbol{P}_{(t=1)} - \boldsymbol{P}_{(t=0)}\right\| < \varepsilon \tag{3}$$

Propensity score matching, as well as other propensity score techniques, are beneficial under conditions where the phenotypic traits and environmental circumstances that influence exposure to the treatment can be directly measured (Guo & Fraser, 2015). However, where some or all of the phenotypic traits and environmental circumstances influencing exposure to the treatment cannot be directly measured, the applicability of propensity score techniques is limited (Guo & Fraser, 2015). Moreover, in addition to the assumptions of a generalized linear regression model (e.g., linearity, normality and heteroscedasticity; Fox, 2016), propensity score matching requires the satisfaction of the assumption of common support and covariate balance (Guo & Fraser, 2015). Common support refers to the assumption that the distribution of the covariates across the treatment and control groups are similar enough to permit matching, while covariate balance refers to the assumption that the similarity in the distribution of the covariates across the treatment and control groups is improved through matching (Guo & Fraser, 2015). These assumptions are fundamental to conducting propensity score matching, as limited common support generates difficulties matching and limited covariate balance reduces the ability to emulate experimental conditions (Guo & Fraser, 2015).

Polygenic Risk Scores

A PRS is an estimate of the effects of multiple alleles on a phenotype, designed to approximate the cumulative genetic likelihood — that is, risk — of an individual experiencing a phenotype. The calculation of a PRS begins with subjecting an independent dataset containing information about the participant's SNPs to a genomewide association study (GWAS; Dudbridge, 2013).[3] A GWAS regresses the phenotype of interest on each SNP using a generalized linear model. The estimated effects of each SNP (β_j) from the independent GWAS is then used to estimate the genetic likelihood of experiencing the phenotype in the dataset of interest. This estimate is created using Eq.4, where the raw polygenic score (PRS_i) is equal to the aggregated value of the coefficient of (β_j) multiplied by the number of reference alleles that individual (i) possesses ($\boldsymbol{G_{ij}}$). The PRS_i can then be transformed to produce the respondents' standardized polygenic scores ($z_{\boldsymbol{PRS}_i}$). Importantly, PRSs are assumed to only represent the additive genetic risk for a phenotype and cannot be used to provide an indication of other genetic processes (e.g., dominant, epistasis and epigenetic risk; Dudbridge, 2013; Euesden et al., 2015). Moreover, PRSs can be calculated using only SNPs that reach a prespecified statistical threshold (typically, $0.05 * 10^{-6}$), or using the whole genome (Francisco & Bustamante, 2018):

$$\boldsymbol{PRS}_i = \sum_{j=1}^{m} \boldsymbol{\beta}_j \boldsymbol{G}_{ij} \tag{4}$$

Given that the estimates from a GWAS are fundamental to the calculation of a PRS, the estimated likelihood is subject to missing heritability. Briefly, while labeled as *missing* heritability, the concept refers to the difference in the variance explained by genetic factors between ACE decomposition models and GWAS. In general, a GWAS produces heritability estimates substantially smaller than ACE decomposition models, suggesting that GWAS cannot explain the variation in a phenotype as well as twin-based models. The missing heritability concept has important implications when discussing PRSs, as the variation in a PRS cannot predict the variation in a phenotype as well as the difference between MZ twins. For instance, a PRS predicts the variance in a phenotype substantially worse than a MZ difference score under conditions where the heritability estimate from a twin study indicates that 60% of the variance in a phenotype is attributed to genetics ($h^2 = .60$), but a GWAS indicates that only 10% of the variance in the phenotype is attributed to genetics ($r^2 = .10$). While suggested to be a function of the number of parameters compared to the sample size, missing heritability has important implications for the GAPS matching approach.

Genetically Adjusted Propensity Score (GAPS) Matching

Similar to propensity and PRS, GAPS matching was designed as a data reduction technique intended to capture respondents'

https://doi.org/10.1017/thg.2022.2 Published online by Cambridge University Press

observed genetic and environmental risk for a phenotypic outcome of interest. A GAPS can be calculated by combining respondents' polygenic scores with respondents' propensity scores and removing any covariation between the terms. The combination of the polygenic score and propensity score can be achieved by introducing the polygenic score (as an independent variable) into the regression analysis used to calculate the propensity score. Although a variety of models can be used to estimate propensity scores (Guo & Fraser, 2015), calculating GAPS for dichotomous constructs provides a straightforward example of the estimation procedure. After calculating and standardizing a PRS, a generalized linear model can be estimated to create a GAPS (Guo & Fraser, 2015). For example, a logistic regression model could be estimated, where the log-odds of a dichotomous variable (identified as $\boldsymbol{v}$) would be regressed on the polygenic score and measures of environmental constructs. This process is captured in Eq. 5, where $\boldsymbol{v}_i$ represents a dichotomously measured outcome variable, $\boldsymbol{Z}_{PRS_i}$ represents a $\mathbf{m} \times \mathbf{n}$ matrix of standardized PRSs (one or more polygenic scores can be included in the estimation procedure), and X_{ki} represents a $\mathbf{m} \times \mathbf{n}$ matrix of environmental constructs:

$$\log\left(\frac{\boldsymbol{v}_i}{1-\boldsymbol{v}_i}\right) = \boldsymbol{\beta}\boldsymbol{Z}_{PRS_i} + \boldsymbol{\beta}X_i \tag{5}$$

After calculating the slope coefficients ($\boldsymbol{\beta}$), Eq. 3 is used to calculate the GAPS. As demonstrated by Eq. 6, GAPS (#xi;$_i$) is equal to the exponential value (*exp*) of a participant's weighted scores on the independent variables ($\boldsymbol{\beta}\boldsymbol{Z}_{PRS_i} + \boldsymbol{\beta}X_i$) divided by 1 + the exponential value of a participant's weighted scores on the independent variables ($\boldsymbol{\beta}\boldsymbol{Z}_{PRS_i} + \boldsymbol{\beta}X_i$). The process of calculating GAPS would be similar if the variable ($\boldsymbol{v}$) was a categorical or continuous variable (Guo and Fraser, 2015):

$$\boldsymbol{\zeta}_i = \frac{\exp(\boldsymbol{\beta}\boldsymbol{Z}_{PRS_i} + \boldsymbol{\beta}X_i)}{1 + \exp(\boldsymbol{\beta}\boldsymbol{Z}_{PRS_i} + \boldsymbol{\beta}X_i)} \tag{6}$$

The result of this estimation process is a single score capturing the observed genetic and environmental risk for a phenotype. The estimation process for calculating GAPS, nonetheless, relies on a series of assumptions that are required to be satisfied. In addition to model assumptions (e.g., linearity, normality and heteroscedasticity; Fox, 2016) and the assumptions associated with PRSs, the estimation of the GAPS requires that the level of measurement for the phenotype remains consistent across all estimations. For example, if the PRS was calculated using a dichotomous measure, the estimation of the GAPS should employ the same dichotomous measure. This requirement is a consequence of the differing amount of variation in a phenotype explained by the PRS and environmental constructs across levels of measurement. As such, if the level of measurement for the dependent construct varies between the models, the variation explained by observed genetic and environmental factors could bias the resulting GAPS.

In addition to the statistical control method or GAPS weighting (Guo & Fraser, 2015), GAPS can be used to match participants to emulate the counterfactual condition created by the MZ twin difference approach. Specifically, respondents with similar observed genetic and observed environmental factors could be matched using the GAPS to mimic the robust counterfactual condition and high internal validity present in a discordant MZ twin model. For instance, respondents with similar GAPS but different educational attainment can be matched, and the resulting matched sample can be used to evaluate the association between educational attainment and future income, which could potentially produce estimates similar to a discordant MZ twin model. Indeed, GAPS can be employed using a variety of matching procedures conventionally used in propensity score matching analyses (see Guo & Fraser, 2015). For instance, GAPS can be used in conjunction with exact matching (Eq. 7) or caliper matching (Eq. 8; ϵ represents the caliper):

$$\boldsymbol{\zeta}_{(t=1)} = \boldsymbol{\zeta}_{(t=0)} \tag{7}$$

$$\left\|\boldsymbol{\zeta}_{(t=1)} - \boldsymbol{\zeta}_{(t=0)}\right\| < \varepsilon \tag{8}$$

Against this backdrop, the current study assessed the validity of the GAPS matching approach by addressing two research questions: (RQ1) What is the degree to which GAPS matching can produce estimates similar to the estimates produced by the MZ twin difference approach? and (RQ2) What is the added value of adjusting for the confounding effects of G × E when estimating the effects of one phenotype on another? To test these research questions, the current study employed simulation analyses to assess how well GAPS matching can approach the estimates produced by the MZ twin difference approach. Nine of 240 simulated comparisons ($N = 10{,}000$ for each simulation) will be discussed in the current study to provide a comprehensive comparison between the GAPS matching and discordant MZ twin methods when estimating associations.[4] The nine simulated examples were selected due to their ability to provide the largest insight into the results across all of the simulated comparisons. Moreover, the selected examples emulate the variance common in complex phenotypes. Specifically, when decomposing the variance in complex phenotypes, a larger portion is attributed to genetic and nonshared environmental factors, while a smaller portion is attributed to the shared environment or unique G × E interactions. That said, all the results for each of the 240 simulations are provided in text files using https://github.com/ianasilver/Genetically-adjusted-propensity-scores-A-comparison-to-discordant-MZ-twin-models,[5] and an R-script for a randomly specified looped simulation is provided as supplemental materials to permit the estimation and evaluation of GAPS matching across randomly specified phenotypes.

Simulation Analyses: Comparing GAPS Matching to the MZ Difference Approach[6]

Assumptions of Simulation Analyses

The simulation analyses discussed below and the corresponding results compare the GAPS matching approach to the discordant MZ twin approach under the optimal conditions for PRSs. These optimal conditions can be highlighted by discussing the three foundational assumptions of the simulation analyses. First, the simulation analyses assume that PRSs can adjust estimates for the heritability of a phenotype in a manner identical to that of the discordant MZ twin approach. Due to the missing heritability associated with PRSs (Manolio et al., 2009), this assumption would likely not hold when employing the GAPS matching approach in a real data example (illustrated below). Moreover, the effects of the missing heritability on the GAPS matching approach are evaluated and discussed below.

Second, the simulation analyses assume that the discordant MZ twin approach does not adjust for variation associated with the

https://doi.org/10.1017/thg.2022.2 Published online by Cambridge University Press

nonshared environment. In a bivariate assessment, this assumption holds true, as the discordance between MZ twins on nonshared environmental conditions will not be adjusted for when estimating the effects of interest (Knopik et al., 2016). Nevertheless, observable or latent measures of the discordance between MZ twins on nonshared environmental conditions can be introduced into multivariate analyses, adjusting for variation in the nonshared environment. While this assumption might not accurately characterize the extent to which the discordant MZ twin approach can generate robust estimates, the simulation analyses provide a comparison to the baseline discordant MZ twin approach. Supplementary material Appendix C, however, provides a replication of the simulation analyses assuming that observable measures of the discordance between MZ twins on nonshared environmental conditions are introduced into a multivariate analysis. As demonstrated, the results suggest that the discordant MZ twin approach possesses an enhanced ability to estimate the true association when observable or latent measures of the discordance between MZ twins on nonshared environmental conditions are introduced into a multivariate analysis.

Finally, due to the difficulties associated with comparing two techniques reliant on different data structures, the simulation analyses assume that the matching error rate will be identical when conducting GAPS matching and difference score analyses using MZ twins. This assumption primarily exists due to the inability to simulate a single dataset that can be representative of *both* singletons and MZ twin pairs, as well as the difficulties associated with ensuring that the effects and error structures are identical across multiple datasets with different simulation specifications (Knopik et al., 2016). The matching error rate of the simulation analyses equally influenced the slope coefficients produced by the MZ difference score evaluation and the GAPS matching approach, which is a limitation of the simulation as MZ twin pairs inherently possess a low matching error rate (i.e., the rate of misidentifying MZ twin pairs; Knopik et al., 2016). Nevertheless, Supplementary Appendix F is provided to demonstrate that the matching procedure itself has limited effects on the observed difference between the discordant MZ twin approach and the GAPS matching approach.

Specification of the Dichotomous Independent Variable (X)

To assess the validity of the GAPS matching approach, 240 unique conditions were simulated to emulate potential phenotypic variation in an independent variable of interest. As demonstrated in Figure 1, the variation in the dichotomous independent variable (***X***) was specified to be equal to a predetermined proportion (*p*) of the genetic (A), shared environment (C), nonshared environment (E) and residual variation (*e*; see mathematical equations in Supplementary Appendix A and available *R*-scripts).[7] Across 240 simulations, three methods were employed to differ the amount of variation contributed to ***X*** by A, C and E. First, variation in ***X*** was equal to .05 incremental increases — from 0 to .95 — in A, while the amount of variation in ***X*** contributed by C and E were equal. The independent variable (***X***) was dichotomized to permit a more intuitive analytical strategy for comparing the GAPS matching approach to the discordant MZ twin approach. To ensure that the estimated scores did not perfectly predict ***X***, the amount of residual variation (*rv*) in ***X*** always equaled .04. For example, the first specification was $\boldsymbol{X} = .000_{\boldsymbol{A}} + .480_{\boldsymbol{E}} + .480_{\boldsymbol{C}} + .040_{\boldsymbol{rv}}$, the midpoint of the distribution of incremental specifications was $\boldsymbol{X} = .450_{\boldsymbol{A}} + .255_{\boldsymbol{E}} + .255_{\boldsymbol{C}} + .040_{\boldsymbol{rv}}$ and the final specification was $\boldsymbol{X} = .950_{\boldsymbol{A}} + .005_{\boldsymbol{E}} + .005_{\boldsymbol{C}} + .040_{\boldsymbol{rv}}$ (Figure 2).

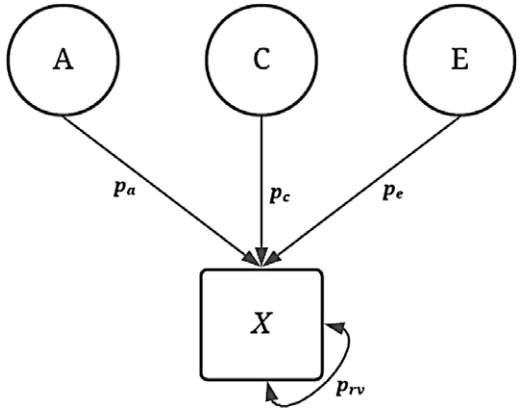

Fig. 1. Visual depiction of specified variation in the simulated treatment variable (***X***) without interactions.
Note: A represents the genetic effect, pa represents the proportion of variation in **X** predicted by A, C represents the shared environment, pc represents the proportion of variation in **X** predicted by C, E represents the nonshared environment, pe represents the proportion of variation in **X** predicted by E, **X** represents a dichotomous treatment construct. The double-headed arrow represents the residual variation in the specification of **X**, where prv represents the proportion of residual variation.

Second, the variation in ***X*** followed the same procedure above (i.e., .05 incremental increases from 0 to .95 for *A*), but the variation in ***X*** contributed by C was three times that contributed by E. For example, the first specification was $\boldsymbol{X} = .000_{\boldsymbol{A}} + .240_{\boldsymbol{E}} + .720_{\boldsymbol{C}} + .040_{\boldsymbol{rv}}$, the middle of the distribution of incremental specifications was $\boldsymbol{X} = .450_{\boldsymbol{A}} + .1275_{\boldsymbol{E}} + .3825_{\boldsymbol{C}} + .040_{\boldsymbol{rv}}$ and the final specification was $\boldsymbol{X} = .950_{\boldsymbol{A}} + .0025_{\boldsymbol{E}} + .0075_{\boldsymbol{C}} + .040_{\boldsymbol{rv}}$. The third condition of variation in ***X*** maintained the same incremental increases in A as the prior conditions, whereas the variation in ***X*** contributed by E was three times that contributed by C (i.e., the reverse of the second condition). For example, the first specification was $\boldsymbol{X} = .000_{\boldsymbol{A}} + .720_{\boldsymbol{E}} + .240_{\boldsymbol{C}} + .040_{\boldsymbol{rv}}$, the middle of the distribution of incremental specifications was $\boldsymbol{X} = .450_{\boldsymbol{A}} + .3825_{\boldsymbol{E}} + .1275_{\boldsymbol{C}} + .040_{\boldsymbol{rv}}$ and the final specification was $\boldsymbol{X} = .950_{\boldsymbol{A}} + .0075_{\boldsymbol{E}} + .0025_{\boldsymbol{C}} + .040_{\boldsymbol{rv}}$.

Subsequently, a series of specifications for ***X*** included a G × E between A and C, A and E, or both. The specifications are illustrated in Figure 3. For each of the specifications including a G × E, 20% of the amount of variation contributed by each component was redirected to the interaction specification (Figure 3). For example, using Panel A of Figure 3, the first specification was $\boldsymbol{X} = .000_{\boldsymbol{A}} + .384_{\boldsymbol{E}} + .480_{\boldsymbol{C}} + .0960_{(\boldsymbol{A}*\boldsymbol{E})} + .040_{\boldsymbol{rv}}$, the middle of the distribution of incremental specifications was $\boldsymbol{X} = .360_{\boldsymbol{A}} + .204_{\boldsymbol{E}} + .255_{\boldsymbol{C}} + .141_{(\boldsymbol{A}*\boldsymbol{E})} + .040_{\boldsymbol{rv}}$ and the final specification was $\boldsymbol{X} = .760_{\boldsymbol{A}} + .004_{\boldsymbol{E}} + .005_{\boldsymbol{C}} + .191_{(\boldsymbol{A}*\boldsymbol{E})} + .040_{\boldsymbol{rv}}$. Using Panel C of Figure 3, the first specification was $\boldsymbol{X} = .000_{\boldsymbol{A}} + .384_{\boldsymbol{E}} + .384_{\boldsymbol{C}} + .096_{(\boldsymbol{A}*\boldsymbol{E})} + .096_{(\boldsymbol{A}*\boldsymbol{C})} + .040_{\boldsymbol{rv}}$, the middle of the distribution of incremental specifications was $\boldsymbol{X} = .270_{\boldsymbol{A}} + .204_{\boldsymbol{E}} + .204_{\boldsymbol{C}} + .141_{(\boldsymbol{A}*\boldsymbol{E})} + .141_{(\boldsymbol{A}*\boldsymbol{C})} + .040_{\boldsymbol{rv}}$ and the final specification was $\boldsymbol{X} = .570_{\boldsymbol{A}} + .004_{\boldsymbol{E}} + .004_{\boldsymbol{C}} + .191_{(\boldsymbol{A}*\boldsymbol{E})} + .191_{(\boldsymbol{a}*\boldsymbol{c})} + .040_{\boldsymbol{rv}}$. Following these examples, 240 specifications of ***X*** were created wherein 60 specifications only included direct effects and 180 specifications included direct and interactive effects. The dichotomous indicator of ***X*** was then used to specify the variation in ***Y***.

https://doi.org/10.1017/thg.2022.2 Published online by Cambridge University Press

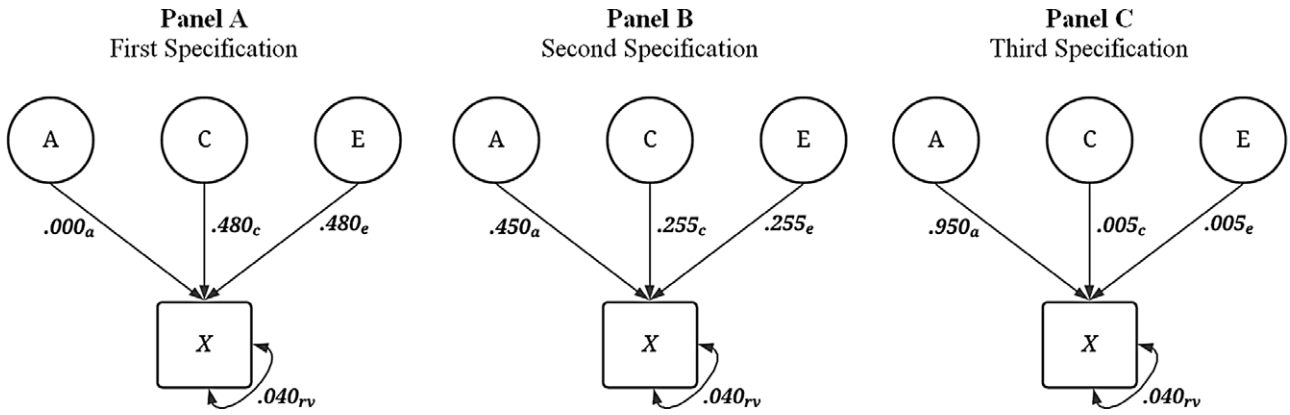

Fig. 2. Visual depiction of specified variation in the simulated treatment variable (***X***) without interactions

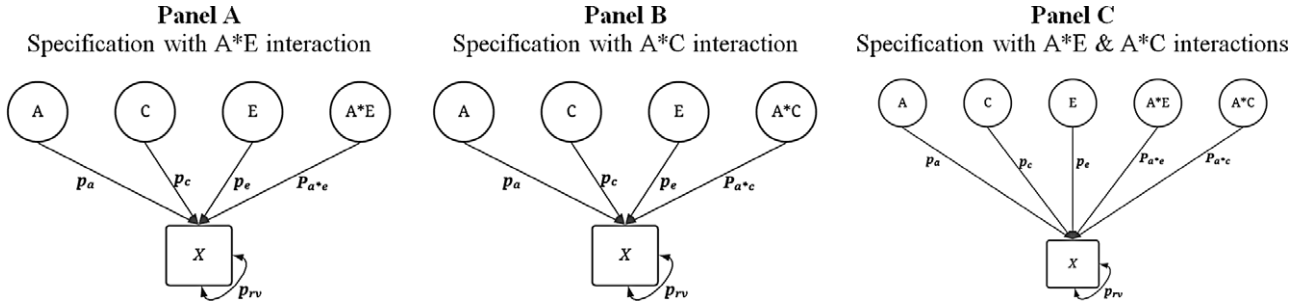

Fig. 3. Visual depiction of specified variation in the simulated treatment variable (***X***) with interactions

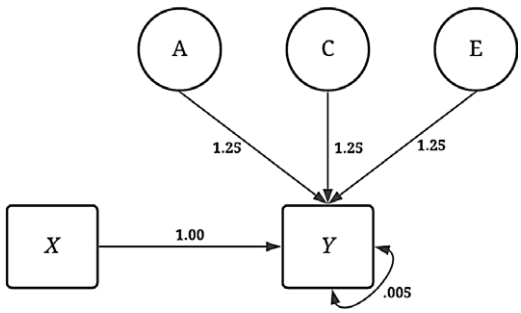

Fig. 4. Visual depiction of specified variation in the simulated independent variable (***Y***) without interactions

Specification of the Dependent Variable (Y)

The simulations were specified in a manner to provide a complete comparison between the estimates produced using GAPS matching to the estimates produced from a discordant MZ twin model. The dependent variable (***Y***) was consistently specified across all the noninteractive conditions of the dichotomous independent variable (***X***). As illustrated by Figure 4, variation in ***Y*** was specified to be the product of the association (i.e., the slope coefficient) between ***X*** and ***Y*** ($b = 1.00$), as well as A (representing the genetic effects), C (representing the shared environment), E (representing the nonshared environment) and residual variation. The influence of A, C and E on ***Y*** was specified as 1.25 to ensure that a substantive amount of variation in ***Y*** was predicted by A, C and E. The specification above generates a substantial amount of confounding in the analytical model and assumes that only a limited amount of error in ***Y*** exists after accounting for the genetic, shared environment and nonshared environment effects.[8]

In addition to the conditions illustrated above, three specifications of the dependent variable with interaction terms were created (see Figure 5) corresponding to the specifications of ***X*** that included the respective interaction. First, an interaction between A and E was specified to predict variation in ***Y***. Second, an interaction between A and C was specified to predict variation in ***Y***. Finally, interactions between A and E as well as A and C were specified to predict variation in ***Y***. Overall, the specifications of the IV and DV provide for a comparison between the GAPS matching and MZ difference score approaches in three conditions: (1) in the absence of interactions, (2) in the presence of a gene-shared environment interaction and (3) in the presence of gene-nonshared environment interactions.

Comparing Discordant MZ twin Approach to GAPS Matching Approach

Although various modeling strategies can be used to compare the discordant MZ twin approach to the GAPS matching approach, the most intuitive procedure is postmatching comparisons. Figure 6 provides visual depictions of the analytical approach. The matching was conducted by regressing ***X*** on all or a proportion of the variation in A, C and E using a binary logistic regression model. The results of the model were then used to calculate a predicted probability — creating a GAPS or a score approximating an MZ difference approach — which determined how cases that experienced the condition ($\boldsymbol{X} = 1$) were matched to cases that did not experience the condition ($\boldsymbol{X} = 0$). When approximating an MZ difference approach, the binary logistic regression model

https://doi.org/10.1017/thg.2022.2 Published online by Cambridge University Press

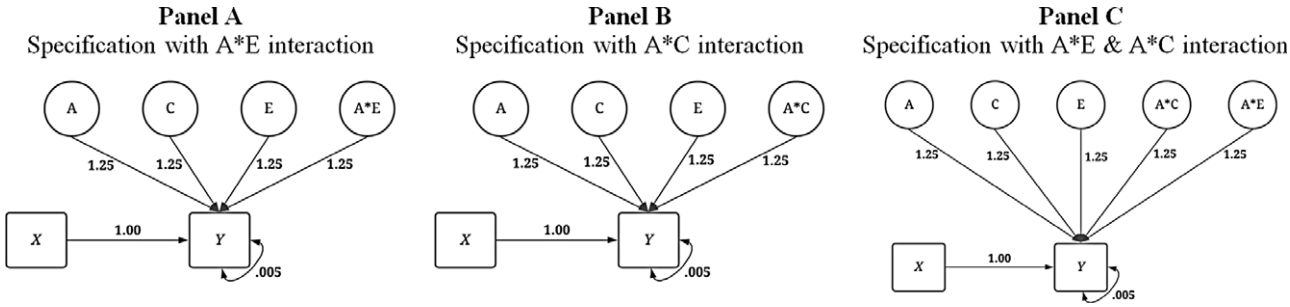

Fig. 5. Visual depiction of specified variation in the simulated independent variable (***Y***) with interactions

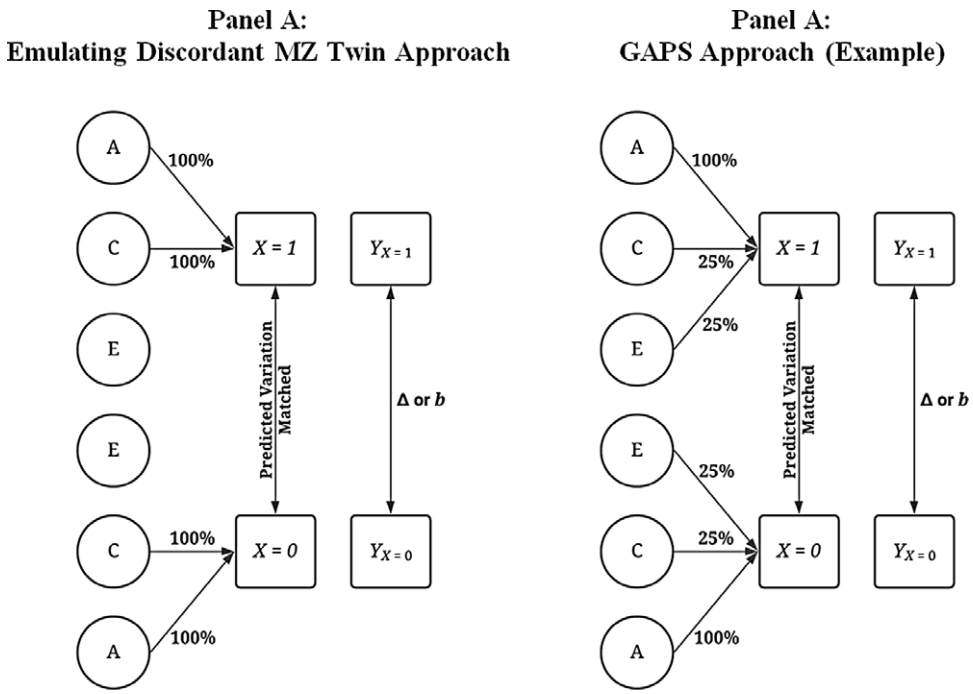

Fig. 6. Visual examples of the analytical matching approach.
Note: Percentage values represent the amount of variation in ***X*** predicted by *A*, *C* and *E* used to predict values on ***X***. For example, .25 of *E* when *E* predicted 38.2% of the variation in *X* indicates that 9.56% of the variation of ***X*** (contributed by the nonshared environment) is adjusted for in the matching process. *A* represents the genetic effect, *C* represents the shared environment, *E* represents the nonshared environment and ***X*** represents a dichotomous independent variable. ***Y*** represents the dependent variable, and Δ or *b* represents the difference in the outcomes between the treatment (***X*** = 1) and control cases (***X*** = 0). The matching procedure described does not fully emulate the discordant MZ twin approach as identical twins would share the same genetic materials. A predicted probability was calculated using a logistic regression with all of or subcomponents of *A* (x1-x4), *C* (x5-x8) or *E* (x9-x12; see Appendix A and R-scripts) serving as the independent variables and ***X*** serving as the dependent variable. The resulting predicted probability was then used to match participants.

regressed ***X*** on all the variation in A and C to emulate the ability to adjust for all of the genetic and shared environmental factors cofounding an association of interest. When estimating the GAPS, ***X*** was regressed on all of the variation in A and a portion of the variation in C and E to emulate the postulated ability to adjust for all of the genetic factors and some of the shared environmental and nonshared environmental factors cofounding an association of interest. Importantly, although the GAPS and score approximating an MZ difference approach were calculated identically, it was assumed that the predicted probability created using all of the variation in A and C would best emulate the MZ difference approach.

Simulated cases that scored a 0 on *X* were matched to simulated cases that scored a 1 on ***X*** using nearest neighbor matching (caliper = .05) at varying amounts of A, C and E predicting scores on ***X***. To emulate the discordant MZ twin approach (Panel A of Figure 6), the amount of information that simulated cases were matched on equaled all of the variation in ***X*** that can be attributed to genetic and shared environmental factors. For instance, if the specification of $\boldsymbol{X} = .450_{\boldsymbol{A}} + .1275_{\boldsymbol{E}} + .3825_{\boldsymbol{C}} + .040_{\boldsymbol{rv}}$, the MZ matching procedure used all the variation in ***X*** contributed by A (.45) and by C (.3825) to match the participants.

While the approach outlined above emulates the discordant MZ twin approach, it is not exact because identical twins would share identical genetic material. The matching technique employed — nearest neighbor matching with a caliper of .05 — only produces pairs with very similar scores on A rather than pairs with identical scores on A (Guo & Fraser, 2015). However, given

https://doi.org/10.1017/thg.2022.2 Published online by Cambridge University Press

that both the MZ twin approach and the GAPS matching approach employed the same seed and a caliper of .05, the differences observed in the slope coefficients using a caliper = .01, .001 or exact matching when comparing the MZ twin approach to the GAPS matching approach were extremely similar to the differences observed when nearest neighbor matching with a caliper of .05 was employed. Nearest neighbor matching was favored in the current simulation due to the substantial amount of time added to the simulation analysis when exact matching was employed, as well as concerns related to the sample size for the postmatching analyses emulating the discordant MZ twin approach. Two *R*-scripts used to conduct the analyses, however, are provided for replication and alteration purposes. After matching the participants, a linear regression model was estimated, where ***Y*** was regressed on ***X*** for only the matched subsample (where A and C were held constant across values of ***X***). The resulting estimate, similar to a discordant MZ twin model, therefore adjusted for the confounding influence of A and C, since shared genetic factors and shared environmental factors were both held constant within the matched subsample.

A similar matching procedure was also employed and cases were matched on their associated GAPS score, rather than on A and C (Panel B of Figure 6). This matching procedure assumed that the all genetic confounding between ***X*** and ***Y*** would be captured by the PRS used to create the GAPS score.[9] In addition to matching participants on all the genetic confounding between ***X*** and ***Y***, the simulated cases were matched to each other using fluctuating amounts of variation contributed to ***X*** by C and E. For example, if the specification of $\boldsymbol{X} = .450_{\boldsymbol{A}} + .1275_{\boldsymbol{E}} + .3825_{\boldsymbol{C}} + .040_{\boldsymbol{rv}}$, one matching procedure would use all the variation contributed by A (.45), 25% of the variation contributed by C (.096) and 25% of the variation from E (.032) to match the simulated cases (Panel B of Figure 6).

Furthermore, interactions between A and E, and/or A and C were included in the matching procedures when the respective specifications of ***X*** and ***Y*** were the focus of the postmatching analysis. In total, for each specification of ***X*** approximately 20 matching procedures were completed, which was designed to provide a comprehensive illustration of how much variation in ***X*** needs to be captured by GAPS to produce estimates similar to estimates produced by the discordant MZ twin approach. All simulated comparisons were completed using the Ohio Supercomputer (Ohio Supercomputer Center, 1987). Below, we highlight specifications of ***X*** and the respective association between ***X*** and ***Y***, to illustrate the information needed for GAPS matching to produce slope coefficients that approach the estimates produced by the discordant MZ twin approach in nine distinct circumstances.

Example 1: No Interactions

Example 1 illustrates the conditions needed for postmatching GAPS slope coefficients to be close to or improve upon the slope coefficients produced by discordant MZ twin methods. In this example, the specification of the variation in ***X*** was set based on a twin-based meta-analysis of common complex traits (Polderman et al., 2015). Specifically, corresponding with complex traits, 45% of the variation in ***X*** was attributed to genetic effects, 38.25% of the variation in ***X*** was attributed to the nonshared environment and 12.75% of the variation in ***X*** was attributed to the shared environment. Therefore, the formula to specify ***X*** was $\boldsymbol{X} = .450_{\boldsymbol{A}} + .3825_{\boldsymbol{E}} + .1275_{\boldsymbol{C}} + .040_{\boldsymbol{rv}}$. The dependent variable (***Y***) was specified following the depiction in Figure 4. Succeeding the matching procedure, the association between ***Y*** and ***X*** was estimated using the matched subsamples and an Ordinary Least Squares (OLS) regression model. The derived estimates and 95% confidence intervals for each postmatching subsample were then plotted to allow for comparison between the MZ and GAPS matching procedures.

Results. Figure 7 provides the estimates and 95% confidence intervals for the association between ***X*** and ***Y*** derived from the matched subsamples. As a reminder, the specified slope coefficient for the association is 1.00 and is represented by the solid vertical line in the figure. The specifications on the *y*-axis provide the proportion of the variation in ***X*** attributed to the specified component used to match participants. For example, 1 * *A* indicates that all the genetic variation in ***X*** (.45) was used to match the participants, and .25 * *E* indicates that 25% of the variation in ***X*** attributed to the nonshared environment (*E*; .3825(.25) = .0955) was used to match the participants. The estimate at the top of Figure 7 (1 * A; 1 * C) provides an approximation for the estimate produced by a discordant MZ twin model. The estimates derived from the matched subsamples, using the indicated matching specifications, in between the upper and lower dashed lines were further from the specified slope coefficient (1.00) than the estimate derived from the approximated MZ twin subsample ($b = 1.253$). The estimates derived from the matched subsamples below the lower dashed line were equal distance or closer to the specified slope coefficient (1.00) than the estimate derived from the approximated MZ twin subsample ($b = 1.253$). Overall, the results suggest that the MZ difference score approached the specified slope coefficient (1.00) more closely than GAPS matching when less environmental information (E and C) is used to create the GAPS. Nevertheless, GAPS matching generally approached the specified slope coefficient (1.00) more closely than discordant MZ twin models when more environmental information (>50% of the variation in E and C) is used to create the GAPS. Overall, the GAPS matching slope coefficient remained substantively different from the specified slope coefficient (1.00) on average ($b \approx 1.25$) across the estimated comparisons. Importantly, these estimates assumed that the polygenic score would capture all the genetic variation in ***X*** (.45).

Examining the effects of missing heritability. To illustrate the effects of missing heritability on estimates derived from GAPS matching, nine additional matching subsamples were created. The specified proportion of the variation in ***X*** attributed to the genetic and environmental components were used to match simulated cases. The results overwhelmingly demonstrated that estimates produced by GAPS matching were further from the specified slope coefficient (1.00) than the discordant MZ twin estimate. This finding suggests that the missing heritability could substantively diminish the ability of GAPS matching to produce estimates equal to the estimates produced by discordant MZ twin methods. Generally, we recognize that GAPS matching will not perfectly emulate the discordant MZ twin approach, because it is unlikely that PRSs will predict the same amount of genetic variation as the discordant MZ twin approach (Table 1; Slatkin, 2009). However, we expect the performance of GAPS matching will be enhanced as larger GWAS are conducted and PRSs are improved.

Results from low and high genetic contribution comparisons. To provide additional comparisons between the GAPS matching approach and discordant MZ twin methods, simulations with low ($\boldsymbol{X} = .100_{\boldsymbol{A}} + .645_{\boldsymbol{E}} + .215_{\boldsymbol{C}} + .040_{\boldsymbol{rv}}$) and

https://doi.org/10.1017/thg.2022.2 Published online by Cambridge University Press

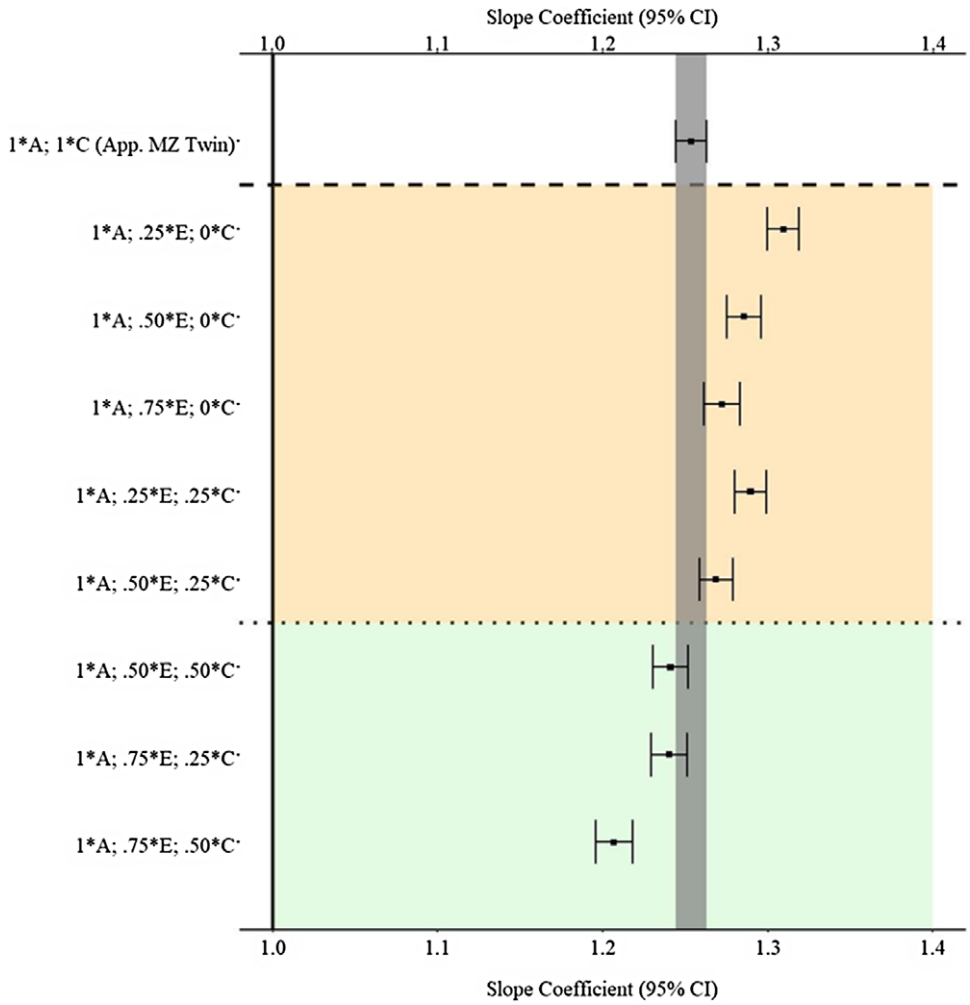

Fig. 7. (Example 1). Slope coefficients of ***Y*** regressed on ***X*** differentially adjusting for the confounding influence of genetic, shared environmental and nonshared environmental effects.

Note: *A* = genetics, *E* = nonshared environment, *C* = shared environment. The true association between ***Y*** and ***X*** is 1.00 (Starting N = 10,000). For the current example, variation in ***X*** was specified as 45 % genetic, 38.2% nonshared environment, 12.8% shared environment and 4% error. Genetic, shared environmental and nonshared environmental factors equally contributed to the variation in ***Y***. All estimates were derived from a postmatching OLS model. Matching was completed using nearest neighbor matching with a caliper of .05. The orange area provides the specifications that performed further away from the true point estimate (1.00) than the approximation of discordant MZ twin methods, and the green area provides the specifications that were closer to the true point estimate (1.00) than the approximation of discordant MZ twin methods. The proportions on the *y*-axis represent the proportion of the variation in ***X*** contributed by the specified component (*A*, *C* or *E*) that is adjusted for by the model. For example, .25 * E indicates that 9.56% of the variation of ***X*** (contributed by the nonshared environment) is adjusted for in the model.

high ($\boldsymbol{X} = .800_{\boldsymbol{A}} + .120_{\boldsymbol{E}} + .040_{\boldsymbol{C}} + .040_{\boldsymbol{rv}}$) genetic contribution to ***X*** were plotted in Figure 8. As demonstrated, the pattern of findings associated with the low and high genetic contribution to ***X*** was similar to the pattern of findings in Figure 7.

Example 2: Interaction Between a and E

Example 2 builds upon the condition presented in Example 1 through the inclusion of a GxE between the genetic effects and the nonshared environment. Distinct from the discordant MZ twin approach, GAPS matching provides the ability to adjust estimates for the confounding effects of genetic-nonshared environment interactions. As such, Example 2 is intended to evaluate if, and how well, the ability to obtain the specified slope coefficient (1.00) is enhanced through the inclusion of a G × E in a GAPS compared to the discordant MZ twin approach. In Example 2, a proportion (~20%) of the variation in ***X*** initially attributed to genetic factors and the nonshared environment was reclassified to be attributed to an interaction between genetic factors and the nonshared environment. As such, the variation in ***X*** was specified as $\boldsymbol{X} = .360_{\boldsymbol{A}} + .306_{\boldsymbol{E}} + .128_{\boldsymbol{C}} + .167_{(\boldsymbol{A}*\boldsymbol{E})} + .040_{\boldsymbol{rv}}$. Additionally, the independent variable (***Y***) for the current example was specified following Figure 5.

Results. Figure 9 provides the estimated slope coefficients of the association between ***X*** and ***Y***, derived from the various matching procedures, for the $A * E$ specification. As demonstrated, the confounding influence of the G × E generates estimates further from the true association (1.00) than the estimates observed in Example 1. Furthermore, distinct from Example 1, only four estimates using the GAPS matching approach were more biased than the estimates derived from the MZ twin approach ($b = 1.405$). The slope coefficients from the discordant MZ twin approach were likely more biased than the majority of the GAPS matching estimates as a result of the inability to capture any variation

https://doi.org/10.1017/thg.2022.2 Published online by Cambridge University Press

Table 1. Demonstration of the effects of missing heritability on Example 1 (simulation starting N = 10,000)

Variation in ***X*** *= A*: 45%; *E*: 38.2%; *C*: 12.8%	Slope coefficient
Specified slope coefficient	1.00
Discordant MZ twin slope coefficient	1.25
Proportion of variation in X	
.25 * *A*; .25 * *E*; .25 * *C*	1.36
.25 * *A*; .50 * *E*; .50 * *C*	1.32
.25 * *A*; .75 * *E*; .75 * *C*	1.29
.50 * *A*; .25 * *E*; .25 * *C*	1.34
.50 * *A*; .50 * *E*; .50 * *C*	1.30
.50 * *A*; .75 * *E*; .75 * *C*	1.25
.75 * *A*; .25 * *E*; .25 * *C*	1.32
.75 * *A*; .50 * *E*; .50 * *C*	1.26
.75 * *A*; .75 * *E*; .75 * *C*	1.21

Note: *A* = genetics, *E* = nonshared environment, *C* = shared environment. For the current example, variation in ***X*** was specified as 45 % genetic, 38.2% nonshared environment, 12.8% shared environment and 4% error. Genetic, shared environmental and nonshared environmental factors equally contributed to the variation in ***Y***. All estimates were derived from a postmatching OLS model. Matching was completed using nearest neighbor matching with a caliper of .05. The proportions represent the proportion of the variation in ***X*** contributed by the specified component (*A*, *C* or *E*) that is adjusted for by the model. For example, .25 * *E* indicates that 9.56% of the variation of ***X*** (contributed by the nonshared environment) is adjusted for in the model.

associated with the nonshared environment or the G × E. Moreover, accounting for the knowledge that missing heritability likely upwardly biases the estimates derived from GAPS matching, the findings suggest GAPS matching can produce estimates close to discordant MZ twin methods if a genetic nonshared environment interaction (G × E) confounds the association between ***X*** and ***Y***.

Results from low and high genetic contribution comparisons. Like Example 1, comparisons with low ($\boldsymbol{X} = .080_{\boldsymbol{A}} + .516_{\boldsymbol{E}} + .215_{\boldsymbol{C}} + .149_{(\boldsymbol{A}*\boldsymbol{E})} + .040_{\boldsymbol{rv}}$) and high ($\boldsymbol{X} = .640_{\boldsymbol{A}} + .096_{\boldsymbol{E}} + .040_{\boldsymbol{C}} + .184_{(\boldsymbol{A}*\boldsymbol{E})} + .040_{\boldsymbol{rv}}$) genetic contribution to ***X*** were also plotted. The findings presented in Figure 10 illustrated that GAPS matching can generally produce slope coefficients that are equivalent or closer to the specified slope coefficient than the discordant MZ twin approach when a G × E confounds the association between ***X*** and ***Y***. At the extreme, Panel B in Figure 10 illustrates that at high genetic contribution to ***X***, GAPS matching can produce estimates similar to or outperform the estimates from the discordant MZ twin approach when any combination of genetic and nonshared environmental information is used to create GAPS. This finding suggests that the ability to adjust estimates for G × E interactions in GAPS matching could be a substantive advancement when compared to the discordant MZ twin approach.

Example 3: Interactions Between a and E, and a and C

The final example, Example 3, demonstrates how estimates using GAPS matching compare the estimates derived from discordant MZ twin methods with two confounding GxE interactions. A proportion of the variation in ***X*** was specified to be attributed to the interactions between genetic factors and the nonshared environment, and genetic factors and the shared environment. The variation in ***X*** for Example 3 was specified as $\boldsymbol{X} = .270_{\boldsymbol{A}} + .306_{\boldsymbol{E}} + .102_{\boldsymbol{C}} + .167_{(\boldsymbol{A}*\boldsymbol{E})} + .116_{(\boldsymbol{A}*\boldsymbol{C})} + .040_{\boldsymbol{rv}}$.

Results. Figure 11 provides the results of the postmatching estimates. Consistent with Example 1, the estimates derived from the approximation of discordant MZ twin approach were closer to the specified slope coefficient than approximately half of the estimates produced with the GAPS matching approach. The divergent findings between Example 2 and Example 3 highlight that discordant MZ twin methods capture (as specified) more variation in ***X*** associated with the shared environment than the nonshared environment. Importantly, the results of Example 3 further demonstrate that the likelihood of producing estimates similar to discordant MZ twin methods increases as more variation in ***X*** is captured within the GAPS. Additionally, capturing variation in ***X*** attributed to G × E can potentially enhance the likelihood of GAPS matching producing slope coefficients similar to discordant MZ twin methods or being closer to the specified slope coefficient than discordant MZ twin methods.

Results from low and high genetic contribution comparisons. In addition to the primary findings of Example 3, comparisons with low ($\boldsymbol{X} = .060_{\boldsymbol{A}} + .516_{\boldsymbol{E}} + .172_{\boldsymbol{C}} + .149_{(\boldsymbol{A}*\boldsymbol{E})} + .063_{(\boldsymbol{A}*\boldsymbol{C})} + .040_{\boldsymbol{rv}}$) and high ($\boldsymbol{X} = .480_{\boldsymbol{A}} + .0096_{\boldsymbol{E}} + .032_{\boldsymbol{C}} + .184_{(\boldsymbol{A}*\boldsymbol{E})} + .168_{(\boldsymbol{A}*\boldsymbol{C})} + .040_{\boldsymbol{rv}}$) genetic contributions to ***X*** are provided in Figure 12. The results illustrated that GAPS matching can produce estimates similar to discordant MZ twin methods when more than 50% of the variation in genetic, shared environmental and nonshared environmental factors is used to create the GAPS. Nevertheless, at high genetic contributions, when G × E exist for both the shared and nonshared environment, the likelihood of GAPS matching producing estimates similar to the discordant MZ twin estimates is diminished. The divergence is likely a function of the increased ability of the discordant MZ twin approach to adjust for interactions between genetic and shared environmental factors.

Real Data Example

Sample

To demonstrate the GAPS matching technique using a real data example, a simple analytical evaluation of the association between completing 4 years of college (Wave 3; 0 = did not complete 4 years of college; 1 = did complete 4 years of college) and personal earnings at Wave 4 (higher values represent higher earnings) was conducted. To conduct this demonstration, the current study relied on the restricted version of the National Longitudinal Survey of Adolescent to Adult Health (Add Health). Briefly, the Add Health is a nationally representative sample of individuals from the USA who were in 7th–12th grade during the 1994–1995 school year. Thus far, there have been five waves of data collection between 1994 and 2019. The Add Health is ideal for the current demonstration as it includes (1) oversampled MZ twin pairs ($N_{\text{pairs}} = 307$; $N = 614$), (2) molecular genetic information from a large portion of the initial sample and (3) a comprehensive survey asking a variety of questions about the shared and nonshared environments. From the initial sample, three subsamples were created. First, a full analytical sample was created using listwise deletion across the two variables of interest ($N = 11{,}587$). Second, a GAPS matched subsample was created using listwise deletion, as well as 10 environmental covariates and 11 PRSs, to match participants that did complete 4 years of college to individuals who did not complete 4 years of college by Wave 3 ($N = 3{,}045$). Finally, an MZ twin subsample was created using listwise deletion

https://doi.org/10.1017/thg.2022.2 Published online by Cambridge University Press

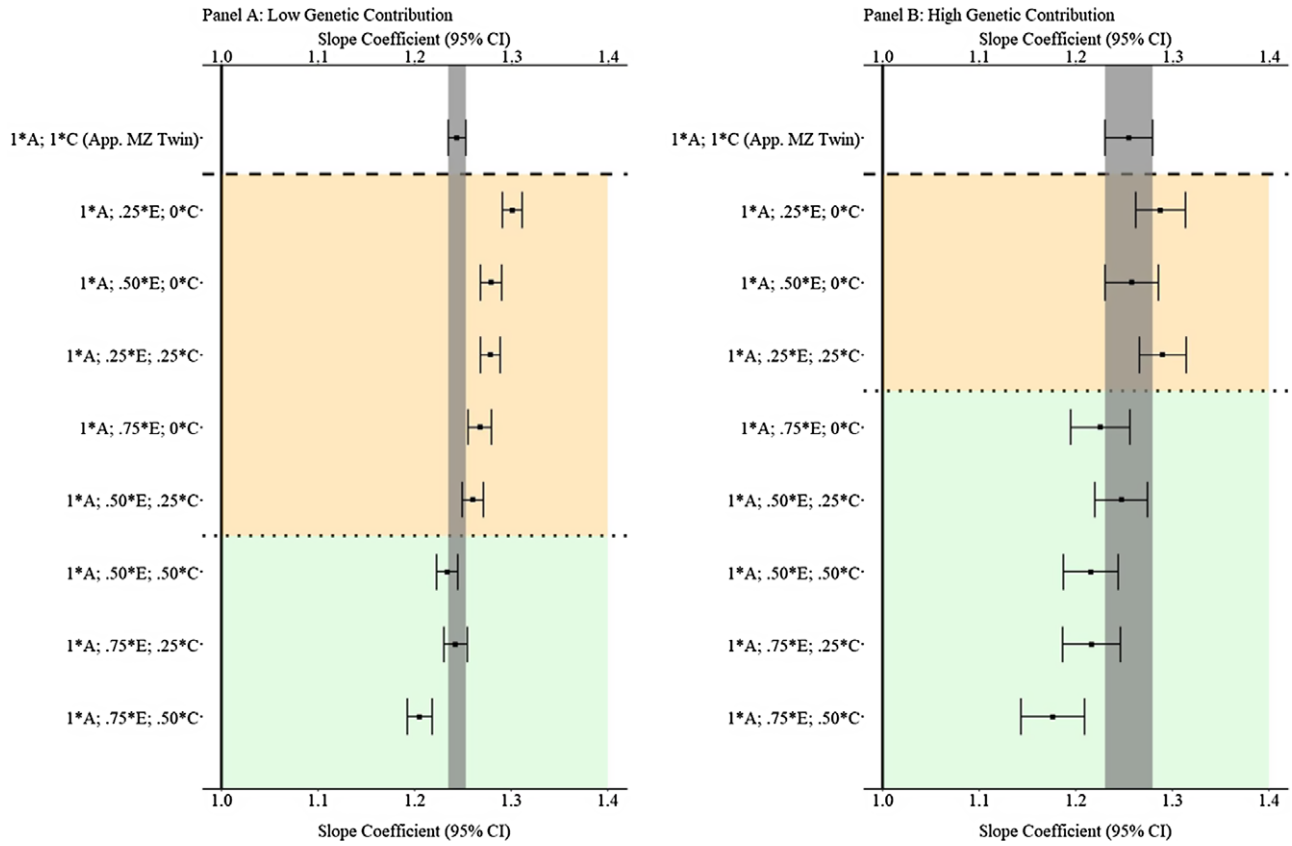

Fig. 8. (Example 1). Slope coefficients of ***Y*** regressed on ***X*** with low and high genetic contribution to the variation in ***X***.
Note: *A* = genetics, *E* = nonshared environment, *C* = shared environment. The true association between ***Y*** and ***X*** is 1.00 (Starting $N = 10{,}000$). For the current example, low genetic contribution: 10 % genetic, 64.5% nonshared environment, 21.5% shared environment and 4% error. High genetic contribution: 80 % genetic, 12% nonshared environment, 4% shared environment and 4% error. Genetic, shared environmental and nonshared environmental factors equally contributed to the variation in ***Y***. All estimates were derived from a postmatching OLS model. Matching was completed using nearest neighbor matching with a caliper of .05. The orange area provides the specifications that performed further away from the true point estimate (1.00) than the approximation of discordant MZ twin methods, and the green area provides the specifications that were closer to the true point estimate (1.00) than the approximation of discordant MZ twin methods. The proportions on the *y*-axis represent the proportion of the variation in ***X*** contributed by the specified component (A, C or E) that is adjusted for by the model.

($N_{pairs} = 189$; $N = 378$). The data cleaning process for all of the measures and the analytical syntax for the current demonstration is provided as a text file titled 'Real data example_GAPS R&R.txt'.

Analytical Strategy

After creating the subsamples, three linear regression models were estimated to evaluate the differential effects of completing 4 years of college (Wave 3) on the log of personal earnings (Wave 4) across the full analytical sample, the GAPS matched subsample and the MZ twin subsample. For the GAPS matched subsample, the matching procedure was conducted by regressing college 4 years on 10 environmental covariates and 11 PRSs to estimate a GAPS, after which, the GAPS was used in conjunction with nearest neighbor matching (caliper = .05) to generate a subsample of participants that completed 4 years of college and participants that did not complete 4 years of college with matching GAPS. Regarding the MZ discordance analysis, a sibling comparison model (using only MZ twins) was estimated permitting the observation of the *between-family effects* of college 4 years on the log of personal income and the *within-family effects* (the genetically sensitive effects) of college 4 years on the log of personal income (Knopik et al., 2016).

Results

Table 2 provides the associations between college 4 years by Wave 3 and the log of personal income at Wave 4 estimated using the full analytical sample, the GAPS matched subsample and the sibling comparison model. As illustrated, completing 4 years of college was positively associated with the log of personal earnings (Wave 4) in the full analytical sample ($b = .539$, $p < .05$). However, the slope coefficient was attenuated to $b = .465$ ($p < .05$) after conducting the GAPS matching procedure. This suggests that adjusting for environmental and genetic factors almost halved the magnitude of the association. In the sibling comparison model, however, the within-family effect was reduced to $b = .459$ and rendered statistically nonsignificant ($p > .05$). Overall, these findings reconfirm the results of the simulation analysis that MZ discordance models will generate more conservative estimates than GAPS matching. Such differences are most likely due to the limitations currently associated with PRSs.

Discussion

Discordant MZ twin methods currently represent one of the foremost statistical tools for producing nongenetically confounded

https://doi.org/10.1017/thg.2022.2 Published online by Cambridge University Press

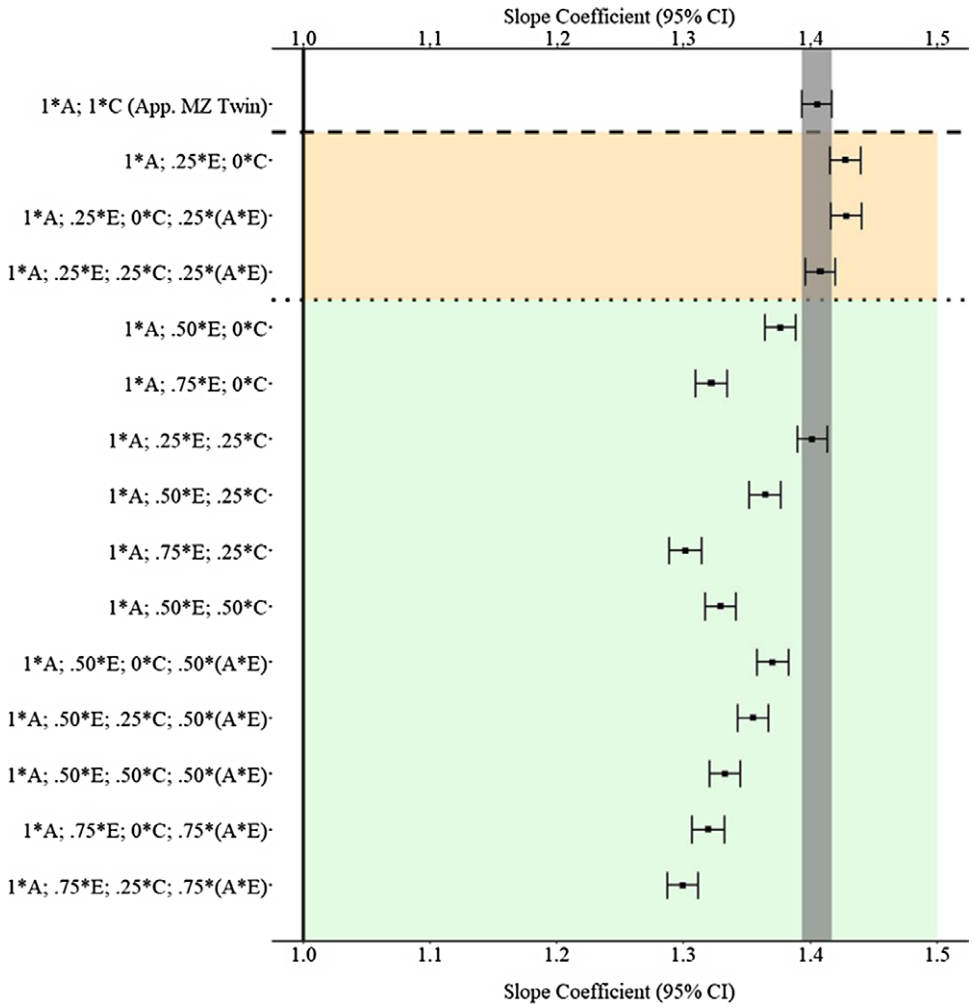

Fig. 9. (Example 2). Slope coefficients of ***Y*** regressed on ***X*** differentially adjusting for the confounding influence of genetic, shared environmental and nonshared environmental effects.

Note: *A* = genetics, *E* = nonshared environment, *C* = shared environment. The true association between ***Y*** and ***X*** is 1.00 (Starting *N* = 10,000). For the current example, variation in ***X*** was specified as 36 % genetic, 30.6% nonshared environment, 12.8% shared environment, 16.7% genetic* nonshared environment and 4% error. Genetic, shared environmental and nonshared environmental factors equally contributed to the variation in *Y*. All estimates were derived from a postmatching OLS model. Matching was completed using nearest neighbor matching with a caliper of .05. The orange area provides the specifications that performed further away from the true point estimate (1.00) than the approximation of discordant MZ twin methods, and the green area provides the specifications that were closer to the true point estimate (1.00) than the approximation of discordant MZ twin methods. The proportions on the *y*-axis represent the proportion of the variation in ***X*** contributed by the specified component (*A*, *C* or *E*) that is adjusted for by the model. For example, .25 * E indicates that 7.7% of the variation of ***X*** (attributed to the nonshared environment) is adjusted for in the model.

estimates in the behavioral sciences (Vitaro et al., 2009). As such, the current study discussed an alternative approach that could potentially produce estimates similar to those of the discordant MZ twin methods. The alternative approach, GAPS matching, integrates variation attributable to environmental risk with variation attributed to genetic risk to adjust estimates for the potential confounding influence of genetic, shared environmental and nonshared environmental factors. The ability of GAPS matching to produce estimates comparable to discordant MZ twin methods was evaluated using 240 simulation analyses. The results from nine of these simulation analyses were reviewed to assess if GAPS matching could approach the estimates produced by discordant MZ twin methods when examining the association between a complex trait (***X***) and an outcome (***Y***). Overall, two findings from the simulated analyses and real data demonstration should be highlighted.

First, when more variation in ***X*** was captured by GAPS, the estimates derived from the postmatching regression models generally appeared to be close to or better than the slope coefficients produced by discordant MZ twin methods. The GAPS matching slope coefficients commonly had 95% confidence intervals that overlapped with the 95% confidence intervals of the slope coefficients produced by discordant MZ twin methods. Moreover, the results also illustrated conditions in which GAPS matching could produce a slope coefficient closer to the specified slope coefficient than the discordant MZ twin approach. This finding suggests that GAPS matching can potentially be a useful technique when discordant MZ twin methodologies cannot be employed to examine the research question of interest (e.g., when genomic data are available, but MZ twins are not).

Second, it appears that the confounding influence of an interaction between genetic and nonshared environmental factors was

https://doi.org/10.1017/thg.2022.2 Published online by Cambridge University Press

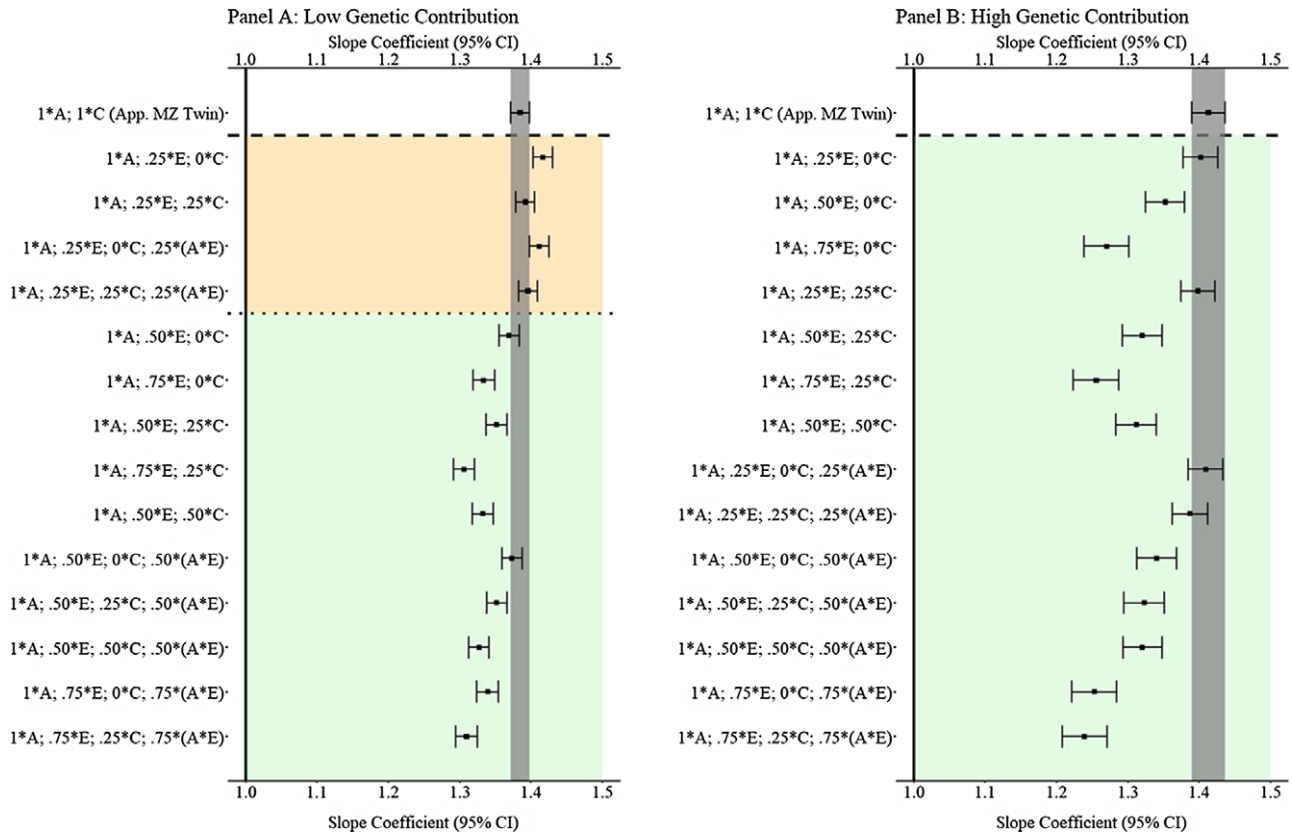

Fig. 10. (Example 2). Slope coefficients of ***Y*** regressed on ***X*** with low and high genetic contribution to the variation in ***X***.
Note: *A* = genetics, *E* = nonshared environment, *C* = shared environment. The true association between ***Y*** and ***X*** is 1.00 (Starting *N* = 10,000). For the current example, Low Genetic Contribution: 8 % genetic, 51.6% nonshared environment, 21.5% shared environment, 14.9% genetic* nonshared environment, and 4% error. High Genetic Contribution: 64 % genetic, 9.6% nonshared environment, 4% shared environment, 18.4% genetic* nonshared environment, and 4% error. Genetic, shared environmental, and nonshared environmental factors equally contributed to the variation in ***Y***. All estimates were derived from a postmatching OLS model. Matching was completed using nearest neighbor matching with a caliper of .05. The orange area provides the specifications that performed further away from the true point estimate (1.00) than the approximation of discordant MZ twin methods, and the green area provides the specifications that were closer to the true point estimate (1.00) than the approximation of discordant MZ twin methods. The proportions on the *y*-axis represent the proportion of the variation in ***X*** contributed by the specified component (*A*, *C*, or *E*) that is adjusted for by the model.

better adjusted for by GAPS matching than discordant MZ twin methods. Given that genetic and nonshared environment interactions contribute to between-individual differences on complex traits in the population (e.g., Purcell, 2002; Knopik et al., 2016), the ability to adjust for G × E represents an important and potentially substantive advancement of GAPS matching when estimating associations between human phenotypes. Through the reliance on PRSs, GAPS matching can provide the opportunity to statistically adjust for confounding variation associated with interactions between genetic and environmental factors in a way that is not possible using the discordant MZ twin approach (e.g., Schmitz & Conley 2017).

Overall, these two key findings demonstrate that the GAPS matching approach can be a promising analytical tool to approximate the estimates produced by discordant MZ twin methodologies. Furthermore, GAPS matching provides the opportunity to test research questions that cannot be assessed using discordant MZ twin methodologies. For instance, GAPS matching can be used to examine the effects of biological sex on psychological outcomes. Additionally, the ability to adjust estimates for G × E is one potential advancement that cannot be achieved when employing discordant MZ twin methodologies. As such, GAPS matching appears to be an analytical strategy that can potentially enhance our understanding of estimated associations.

Limitations of GAPS Matching

Limitations accompanying polygenic scores can diminish the ability to estimate a true GAPS (i.e., their genetic and environmental likelihood). The most influential limitation is the *missing heritability* problem (Slatkin, 2009). As evident by heritability estimates derived from GWAS and the real data evaluation conducted in the current study, polygenic scores rarely explain the amount of variation in a phenotype projected by twin-based heritability studies (Manolio et al., 2009). While various factors likely contribute to missing heritability (e.g., Zuk et al., 2012; Zuk et al., 2014), scholars often suggest that the missing heritability is equated to the inability to achieve enough statistical power (Bloom et al., 2013; van der Sluis et al., 2010). As demonstrated in Example 1, the missing heritability in polygenic scores limits the capacity of GAPS to capture genetic confounding as well as the MZ difference approach.

In addition to missing heritability, PRSs only capture the effects of common variants on a phenotype and do not capture variation

https://doi.org/10.1017/thg.2022.2 Published online by Cambridge University Press

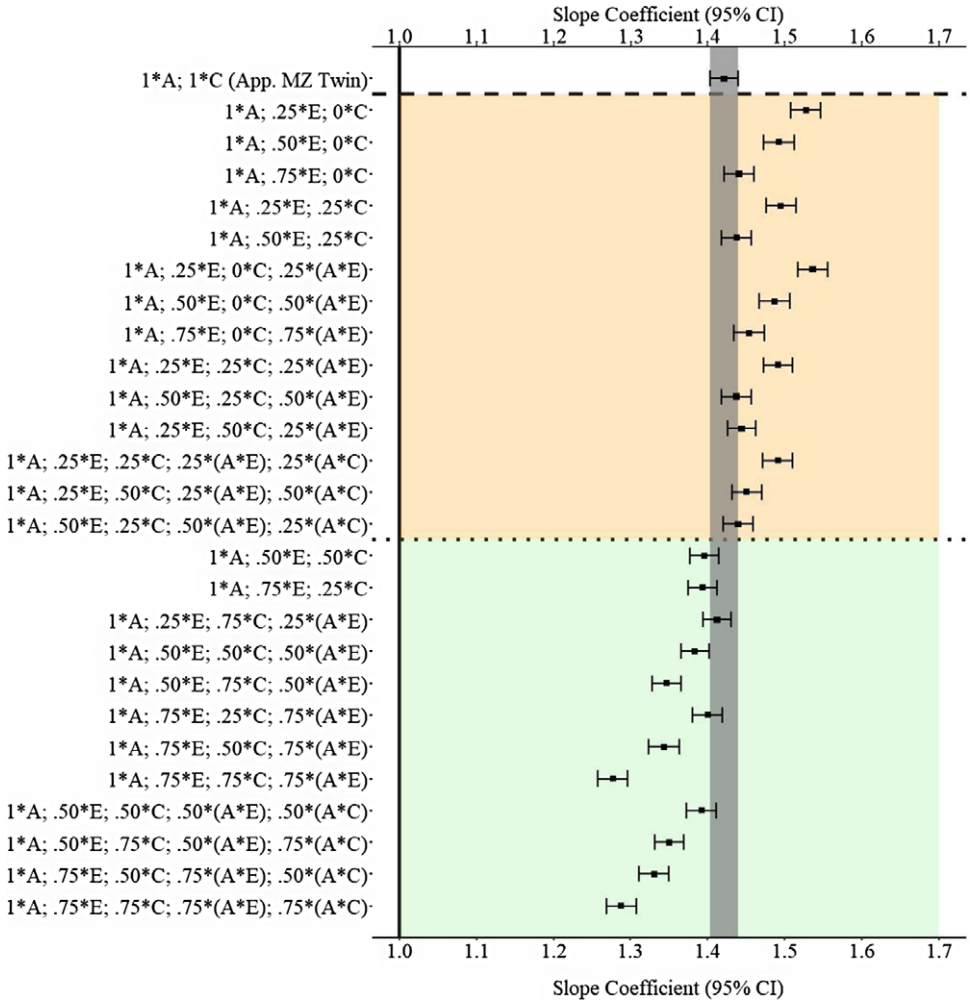

Fig. 11. (Example 3). Slope coefficients of ***Y*** regressed on ***X*** differentially adjusting for the confounding influence of genetic, shared environmental and nonshared environmental effects.

Note: *A* = genetics, *E* = nonshared environment, *C* = shared environment. The true association between ***Y*** and ***X*** is 1.00 (Starting *N* = 10,000). For the current example, variation in ***X*** was specified as 27 % genetic, 30.6% nonshared environment, 10.2% shared environment, 16.7% genetic* nonshared environment, 11.6% genetic*shared environment and 4% error. Genetic, shared environmental and nonshared environmental factors equally contributed to the variation in ***Y***. All estimates were derived from a postmatching OLS model. Matching was completed using nearest neighbor matching with a caliper of .05. The orange area provides the specifications that performed further away from the true point estimate (1.00) than the approximation of discordant MZ twin methods, and the green area provides the specifications that were closer to the true point estimate (1.00) than the approximation of discordant MZ twin methods. The proportions on the *y*-axis represent the proportion of the variation in ***X*** contributed by the specified component (*A*, *C* or *E*) that is adjusted for by the model. For example, .25 * E indicates that 7.7% of the variation of ***X*** (attributed to the nonshared environment) is adjusted for in the model.

related to the mechanisms influencing specific genetic effects (Purcell et al., 2009). As such, the inclusion of a PRS does not permit GAPS matching to adjust estimates for the influence of rare variants, structural variation and epigenetics on the association between a phenotype and an outcome (Dudbridge, 2013; Wray et al., 2014). Although the effects of missing heritability and other limitations of PRSs on GAPS matching appears to diminish the value of the technique, other research teams have begun to produce scores with heritability estimates approaching estimates produced from twin-based heritability studies (Young, 2019). Moreover, scholars are currently developing strategies to capture variation pertaining to the mechanisms causing specific genetic effects when calculating PRSs, including the identification of rare variants (Crouch & Bodmer, 2020) and biologically informed polygenic scores (Hari Dass et al., 2019). As such, if the analytical sample and techniques for GWAS and polygenic scores continue to advance, the effects of missing heritability and the inability to quantify the mechanisms causing specific genetic effects on the estimates produced by GAPS matching should be diminished (Young, 2019).

Limitations of the Simulation Analyses

In addition to the limitations associated with GAPS matching, the current simulation was tasked with comparing a twin-based statistical methodology to a methodology designed for singletons. As such, while matching represented the foremost strategy to potentially compare the discordant MZ twin approach to the GAPS matching approach, the simulation analysis was limited in developing matched 'twin' pairs — pairs matched using

https://doi.org/10.1017/thg.2022.2 Published online by Cambridge University Press

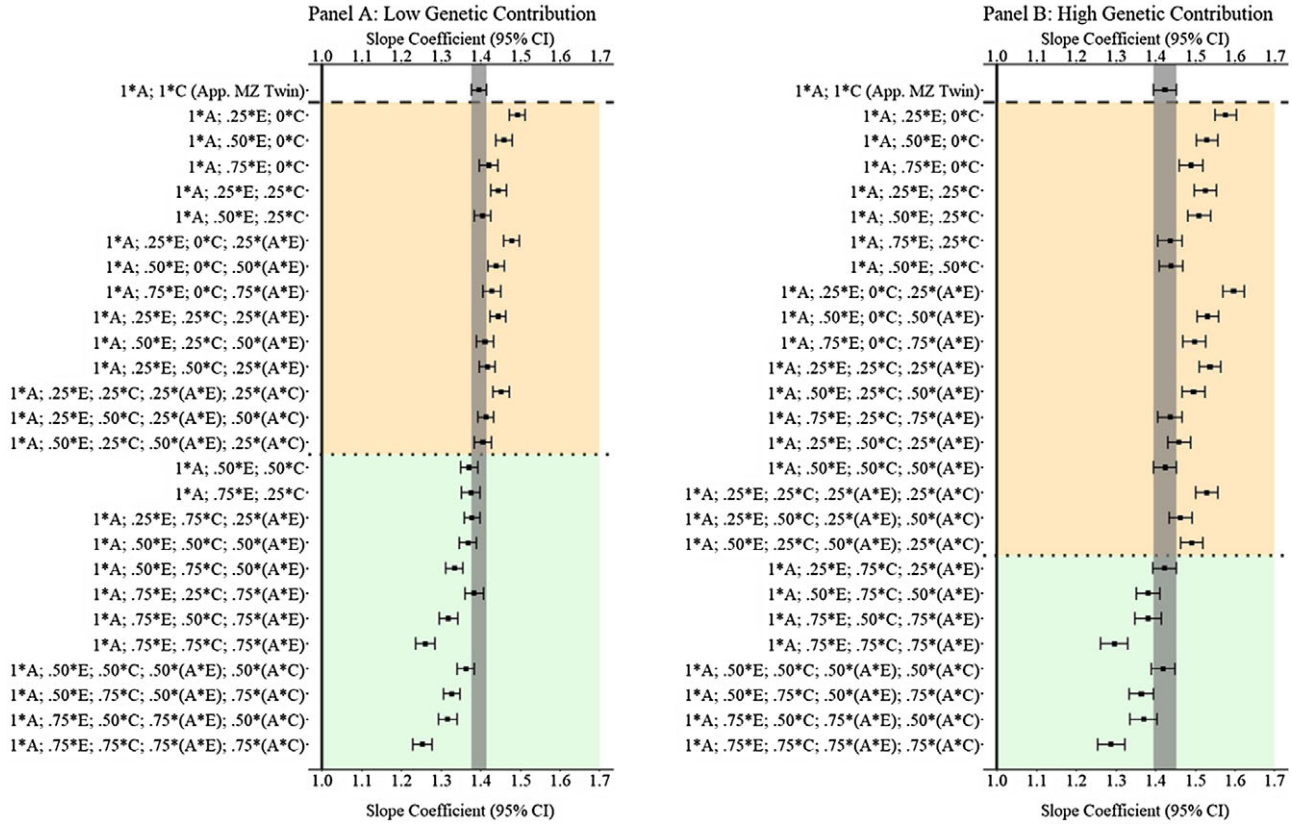

Fig. 12. (Example 3). Slope coefficients of ***Y*** regressed on ***X*** with low and high genetic contribution to the variation in ***X***.
Note: *A* = genetics, *E* = nonshared environment, *C* = shared environment. The true association between ***Y*** and ***X*** is 1.00 (Starting N = 10,000). For the current example, Low genetic contribution: 6% genetic, 51.6% nonshared environment, 17.2% shared environment, 14.9% genetic* nonshared environment, 6.3% genetic*shared environment and 4% error. High genetic contribution: 48% genetic, .96% nonshared environment, 3.2% shared environment, 18.4% genetic* nonshared environment, 16.8% genetic*shared environment and 4% error. Genetic, shared environmental and nonshared environmental factors equally contributed to the variation in ***Y***. All estimates were derived from a postmatching OLS model. Matching was completed using nearest neighbor matching with a caliper of .05. The orange area provides the specifications that performed further away from the true point estimate (1.00) than the approximation of discordant MZ twin methods, and the green area provides the specifications that were closer to the true point estimate (1.00) than the approximation of discordant MZ twin methods. The proportions on the *y*-axis represent the proportion of the variation in ***X*** contributed by the specified component (*A*, *C* or *E*) that is adjusted for by the model.

A and C — that were identical. As such, future scholarship should employ real twin data with genomic data to compare the methodologies and provide an approximation of the bias present in the GAPS approach due to the limitations associated with polygenic scores. Moreover, the comparisons on the simulated datasets conducted using the 95% confidence intervals are subject to the limitations associated with frequentist probability theory (Fox, 2016; Law, 2015). As such, we believe it would be unwise to compare these approaches using the 95% confidence intervals when applying the approach to real data given that twin samples and samples of singletons with molecular genetic information can be structurally different. Additionally, due to difficulties specifying a simulation where the resulting dataset is representative of *both* twins *and* singletons, the current study relied on the assumption that the matching error rate would be identical when conducting GAPS matching and difference score analyses using MZ twins. Due to this assumption, the estimates resulting from the discordant MZ twin approach are likely upwardly biased because of the relatively low rate of misidentification associated with MZ twin pairs (Knopik et al., 2016). As such, future research should consider this limitation when implementing the GAPS matching approach in an effort to more closely emulate discordant MZ twin models.

Conclusion

Long considered the foremost statistical approach to producing nongenetically confounded estimates, discordant MZ twin methodologies provide a unique opportunity to adjust estimates for the confounding influence of genetic and shared environmental factors. The simulation analyses conducted in the current study illustrated the robust nature and validity of discordant MZ twin methodologies. Challenges associated with the discordant MZ twin approach, however, remain. Considering these challenges, the GAPS matching approach was discussed, and simulation and real data analyses were conducted to compare the estimates produced using the GAPS matching approach to the estimates produced using the discordant MZ twin approach. While the simulations illustrated favorable findings, current challenges associated with GAPS matching — illustrated in the real data example — hinder the ability for users to estimate slope coefficients similar to the slope coefficients produced using discordant MZ twin methodologies in all conditions. Nevertheless, as missing heritability diminishes and rich survey data is collected alongside whole genome data, GAPS matching could present a unique opportunity to approximate or advance upon discordant MZ twin methodologies.

https://doi.org/10.1017/thg.2022.2 Published online by Cambridge University Press

Table 2. Estimating the association between completing 4 years of college and personal earnings using the full Add Health sample, the GAPS matched sample and a sibling comparison model

	DV: Log personal earnings (Wave 4)					
	Full analytical sample		GAPS matched sample		Sibling comparison model	
	b	*se*	*b*	*se*	*b*	*se*
College 4 years	0.539*	0.026	0.465*	0.046	0.606*	0.131
College 4 years (*within-family effects)*	–	–	–	–	−0.459	0.323
N	11,587		3,045		378	
GAPS matching DV: college 4 years						
			b	*se*		
Environmental covariates (Wave 1)						
Parental separation			−0.152	0.576		
Parents worked			0.173	0.100		
Cognitive abilities			0.081*	0.006		
Household income			0.002*	0.001		
Maternal education			0.120*	0.058		
Paternal education			0.335*	0.058		
Biological sex			−0.424*	0.153		
Ancestry			0.140*	0.053		
Smoked cigarettes			−0.124	0.097		
Major injury			−0.363	0.268		
Polygenic risk scores						
Body mass index			−0.127*	0.051		
Height			−0.089	0.053		
Cigarettes smoked per day			−0.074	0.049		
Extraversion			0.073	0.049		
Attention-deficit hyperactivity disorder			−0.099*	0.050		
Bipolar disorder			−0.121*	0.060		
Major depressive disorder			−0.073	0.058		
Schizophrenia			0.087	0.086		
Mental health cross disorder			0.009	0.066		
Alzheimer			0.069	0.051		
Educational attainment			0.155*	0.053		
N			3,621			

Note: GAPS matching was completed by using a logistic regression model of college 4 years regressed on the specified covariates to estimate a GAPS (predicted probability). After estimating the predicted probability, nearest neighbor matching with a caliper of .05 was completed to create the GAPS matched subsample of 3,315 participants. Listwise deletion was used to create the analytical samples for the current example.
*$p < .05$.

Acknowledgments. The authors want to thank J.C. Barnes for his insight into conducting the simulation analyses.

Notes

1 A similar method was employed by Schmitz and Conley (2016) to assess the impact of job loss on BMI.
2 In the context of propensity score matching, 'treatment' identifies the cases exposed to the condition of interest, while 'control' identifies cases that represent an appropriate counterfactual condition (Guo & Fraser, 2015).
3 The employment of an independent dataset is required to reduce the influence of sample bias on the estimated polygenic risk scores (Dudbridge, 2013).
4 The mathematical formulas for the data specifications are provided in Appendix A. All the code for the current study is provided as supplemental material. Due to the employment of simulation analysis, the current study was exempt from acquiring IRB approval.
5 Brief directions: copy link into web browser and download the zip folder titled 'Complete Results.zip', which contains all the text files (.txt) corresponding to the 240 simulations labeled by the amount of variation in each treatment (***X***) caused by A, C and E.
6 Appendix B and Appendices C–G provide a description and the results (respectively) of various supplemental and sensitivity analyses conducted to further evaluate the GAPS matching approach.
7 Residual variation 'rv' was included in the model to ensure that the subsequent regression models did not perfectly predict *X* or *Y*. *X* was initially simulated as a continuous construct ranging in values between 0 and 1. After simulating *X* as a continuous construct that ranged between 0 and 1, *X* was dichotomized where scores above .50 received a value of 1 and scores equal to or below .50 received a value of 0.

https://doi.org/10.1017/thg.2022.2 Published online by Cambridge University Press

8 Considering the sources of error, after accounting for the genetic, shared environment and nonshared environment effects (*E*) effects, only measurement error and random error can exist in *Y*.

9 To demonstrate the effects of missing heritability in the polygenic score, study 1 (introduced below) was re-estimated to illustrate the effects on GAPS when the polygenic scores only captured a limited amount of the variation in *X* contributed by *A*.

References

Asbury, K., Dunn, J. F., Pike, A., & Plomin, R. (2003). Nonshared environmental influences on individual differences in early behavioral development: a monozygotic twin differences study. *Child Development*, *74*, 933–943.

Bloom, J. S., Ehrenreich, I. M., Loo, W. T., Lite, T. L. V., & Kruglyak, L. (2013). Finding the sources of missing heritability in a yeast cross. *Nature*, *494*, 234–237.

Crouch, D. J., & Bodmer, W. F. (2020). Polygenic inheritance, GWAS, polygenic risk scores, and the search for functional variants. *Proceedings of the National Academy of Sciences*, *117*, 18924–18933.

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS Genetics*, *9*, e100334.

Euesden, J., Lewis, C. M., & O'Reilly, P. F. (2015). PRSice: Polygenic risk score software. *Bioinformatics*, *31*, 1466–1468.

Fox, J. (2016). *Applied regression analysis, linear models, and related methods*. Sage Publications.

Francisco, M., & Bustamante, C. D. (2018). Polygenic risk scores: A biased prediction? *Genome Medicine*, *10*, 1–3.

Guo, S., & Fraser, M. W. (2015). *Propensity score analysis* (vol. *12*). Sage Publishing.

Hari Dass, S. A., McCracken, K., Pokhvisneva, I., Chen, L. M., Garg, E., Nguyen, T. T., Wang Z, Barth, B., Yaqubi, M., McEwen, L. M., MacIsaac, J. L., Diorio, J., Kobor, M. S., O'Donnell, K. J., Meaney, M. J., & Silveira, P. P. (2019). A biologically-informed polygenic score identifies endophenotypes and clinical conditions associated with the insulin receptor function on specific brain regions. *EBioMedicine*, *42*, 188–202.

Knopik, V. S., Ndiderhiser, J. M., DeFries, J. C., & Plomin, R. (2016). *Behavioral Genetics* (7th ed.). Worth Publishers.

Law, A. M. (2015). *Simulation modeling and analysis* (5th ed.). McGraw-Hill.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., . . . **Visscher, P. M.** (2009). Finding the missing heritability of complex diseases. *Nature*, *461*, 747–753.

Motz, R. T., Barnes, J. C., Caspi, A., Arseneault, L., Cullen, F. T., Houts, R., Wertz, J., & Moffitt, T. E. (2019). Does contact with the justice system deter or promote future delinquency? Results from a longitudinal study of British adolescent twins. *Criminology*, *58*, 307–335.

Ohio Supercomputer Center. (1987). Ohio Supercomputer Center. Columbus OH: Ohio Supercomputer Center. http://osc.edu/ark:/19495/f5s1ph73.

Oskarsson, S., Thisted Dinesen, P., Dawes, C. T., Johannesson, M., & Magnusson, P. K. E. (2017). Education and social trust: testing a causal hypothesis using the discordant twin design. *Political Psychology*, *38*, 515–531.

Polderman, T. J., Benyamin, B., De Leeuw, C. A., Sullivan, P. F., Van Bochoven, A., Visscher, P. M., & Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics*, *47*, 702–709.

Purcell, S. (2002). Variance components models for gene–environment interaction in twin analysis. *Twin Research and Human Genetics*, *5*, 554–571.

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F. & Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, *460*, 748–752.

Ross, J. M., Ellingson, J. M., Rhee, S. H., Hewitt, J. K., Corley, R. P., Lessem, J. M., & Friedman, N. P. (2020). Investigating the causal effect of cannabis use on cognitive function with a quasi-experimental co-twin design. *Drug and Alcohol Dependence*, *206*, 107712.

Schmitz, L. L., & Conley, D. (2016). *The impact of late-career job loss and genotype on Body Mass Index* (No. w22348). National Bureau of Economic Research.

Schmitz, L. L., & Conley, D. (2017). Modeling gene-environment interactions with quasi-natural experiments. *Journal of Personality*, *85*, 10–21.

Silberg, J. L., Copeland, W., Linker, J., Moore, A. A., Roberson-Nay, R., & York, T. P. (2016). Psychiatric outcomes of bullying victimization: A study of discordant monozygotic twins. *Psychological Medicine*, *46*, 1875–1883.

Slatkin, M. (2009). Epigenetic inheritance and the missing heritability problem. *Genetics*, *182*, 845–850.

Thornton, L. M., Trace, S. E., Brownley, K. A., Ålgars, M., Mazzeo, S. E., Bergin, J. E., Maxwell, M., Lichtenstein, P., Pedersen, N. L., & Bulik, C. M. (2017). A comparison of personality, life events, comorbidity, and health in monozygotic twins discordant for anorexia nervosa. *Twin Research and Human Genetics*, *20*, 310–318.

Tiu, R. D., Wadsworth, S. J., Olson, R. K., & DeFries, J. C. (2004). Causal models of reading disability: A twin study. *Twin Research and Human Genetics*, *7*, 275–283.

Van Der Sluis, S., Verhage, M., Posthuma, D., & Dolan, C. V. (2010). Phenotypic complexity, measurement bias, and poor phenotypic resolution contribute to the missing heritability problem in genetic association studies. *PLoS One*, *5*, e13929.

Vitaro, F., Brendgen, M., & Arseneault, L. (2009). The discordant MZ-twin method: One step closer to the holy grail of causality. *International Journal of Behavioral Development*, *33*, 376–382.

Wray, N. R., Lee, S. H., Mehta, D., Vinkhuyzen, A. A., Dudbridge, F., & Middeldorp, C. M. (2014). Research review: Polygenic methods and their application to psychiatric traits. *Journal of Child Psychology and Psychiatry*, *55*, 1068–1087.

Young, A. I. (2019). Solving the missing heritability problem. *PLoS Genetics*, *15*, e1008222.

Zuk, O., Hechter, E., Sunyaev, S. R., & Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, *109*, 1193–1198.

Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M. J., Neale, B. M., Sunyaev, S. R., & Lander, E. S. (2014). Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences*, *111*, E455–E464.

https://doi.org/10.1017/thg.2022.2 Published online by Cambridge University Press