

The use of polygenic risk scores as a covariate in psychological studies

Md Jamal Uddin^{a,b,d,e,f,*}, Carsten Hjorthøj^{b,c,d}, Tanvir Ahammed^e, Merete Nordentoft^{b,d}, Claus Thorn Ekstrøm^a

^a Section of Biostatistics, Department of Public Health, University of Copenhagen, Denmark

^b Copenhagen University Hospital, Mental Health Centre Copenhagen, Denmark

^c Section of Epidemiology, Department of Public Health, University of Copenhagen, Denmark

^d The Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH), Copenhagen and Aarhus, Denmark

^e Department of Statistics, Shahjalal University of Science and Technology, Sylhet, Bangladesh

^f Department of General Educational and Development, Daffodil International University, Dhaka, Bangladesh

ARTICLE INFO

Keywords:

Genetic risk scores/polygenic risk scores

Causal association

Covariate

Simulation

Bias

ABSTRACT

Polygenic risk scores (PRS) are increasingly used to predict diseases (e.g., schizophrenia). However, the proper application of the PRS in psychological studies is sparse. We aimed to evaluate the methodological aspects of using PRS as a covariate in causal association studies in psychology. We conducted a simulation study using realistic scenarios, such as gene and exposure or confounder interaction, gene and gene interaction, gene acts as confounding and instrumental variable and replicated each scenario 1000 times. We found that when the genotype(s) that was used for calculating PRS interacted with the exposure and if the analyses considering this interaction, the exposure effect similar with the true value. However, a significant amount of bias was present even after adjustment for the measured confounders i) when the gene interacted with the unmeasured confounder and this bias amplified when the interaction effect was two times stronger; and ii) when the gene acts as a confounder or instrumental variable. Therefore, theoretical knowledge of psychology should be incorporated when the PRS is utilized in psychological causal association studies in order to reduce systematic bias and improve the precision of the exposure effect.

1. Background

In psychological studies, the assessment of the causal association between exposure and disease generally involves modelling the environmental and genetic causes of disease. For example, environmental factors such as location and season of birth, maternal obstetrical complications, parental age, and low social class may cause mental disorders (e.g., schizophrenia) (Gattaz et al., 2002; Marangoni et al., 2018; Pries et al., 2018; Richardson et al., 2018; Vilain et al., 2013), and the genetic contribution to the etiology of several diseases (e.g., schizophrenia, bipolar disorder) is well documented from genome-wide association studies (GWAS) (Anderson et al., 2019; Benca et al., 2017; Mistry et al., 2018b; Power et al., 2015; Ripke et al., 2014). For instance, 108 genetic loci have been identified for schizophrenia (Ripke et al., 2014).

As GWAS have become increasingly popular due to their decreasing cost and stable nature, individual genetic risk scores are often constructed by combining the effects of genetic variants into a single measure that summarizes the combined evidence for a genomic burden for

diseases or predicts a patient's response drug treatment or adverse drug reactions (Clare et al., 2017; Lally and MacCabe, 2016; Ripke et al., 2014; Visscher et al., 2017). In particular, the polygenic risk score (PRS) is widely used for predicting diseases and to adjust for the genetic background without specifying a detailed genetic model (Agerbo et al., 2015; Euesden et al., 2015; Ripke et al., 2014). The PRS has a lot of promise to help with the etiology of psychopathology and the individual risk measurement, which together might help with the diagnosis and cure of psychiatric disease (Bogdan et al., 2018). Several previous studies (Cecile et al., 2011; Dima and Breen, 2015; Dudbridge, 2013, 2016; Escott-Price et al., 2017; Lewis and Vassos, 2017; Maher, 2015; Mistry et al., 2018a; Ott, 2015; Vilhjálmsdóttir et al., 2015; Walter et al., 2013; Wray et al., 2014) demonstrated the methodological and clinical aspects of the PRS models in GWAS. For example, Dudbridge (Dudbridge, 2013, 2016) and Wray et al. (2014) explained and reviewed the power and predictive accuracy, variance estimation, and sample size issues of the PRS; Dima and Breen (2015) described the usefulness and application, including clinical relevance of the PRS for schizophrenia

* Corresponding author. Department of Statistics, Shahjalal University of Science and Technology, Sylhet, Bangladesh.

E-mail address: jamal-sta@sust.edu (M.J. Uddin).

<https://doi.org/10.1016/j.metip.2022.100099>

Received 29 December 2021; Received in revised form 27 June 2022; Accepted 30 June 2022

Available online 8 July 2022

2590-2601/© 2022 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

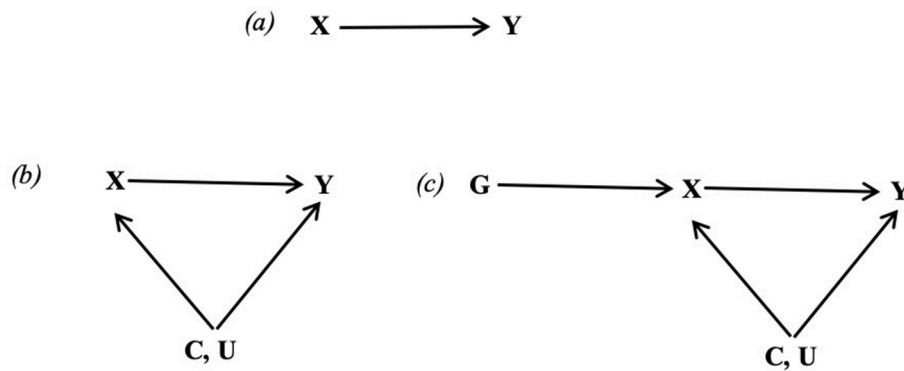


Fig. 1. Demonstration of the Directed acyclic graph, concept of confounding and instrumental variable. X = exposure, Y = outcome, C = observed confounder and U = unobserved confounder, G = genotype (instrumental variable).

and bipolar disorder; So and Sham (So and Sham, 2017) presented empirical Bayesian approaches to improve polygenic risk prediction from summary statistics; and Chun et al. (2018) proposed non-parametric partitioning-based risk prediction model using GWAS summary statistics.

While it is enticing to include the PRS as a predictor or a variable that summarizes the genetic information in a model, the usefulness of the PRS as a covariate, including gene-gene or gene-exposure or gene-confounder interactions on the causal association between an exposure and a disease or other outcomes is not well recognized or unclear in the literature (Agerbo et al., 2004; Benros et al., 2016; Compton and Manseau, 2017; Manrique-Garcia et al., 2014). For example, when studying the effect of cannabis use on the risk of schizophrenia and researchers want to adjust the PRS for schizophrenia in this association model, the effect of cannabis may be questionable if the genetic marker (s) that are used for calculating the PRS act as confounding or there is a gene-cannabis interaction (Colizzi et al., 2015; Klaus et al., 2017; Trotta et al., 2016). Another example, if the researcher wants to investigate the possible effect of the PRS for schizophrenia on the association between infections and the risk of schizophrenia, the PRS-infections interaction may induce bias in the association unless the interaction is considered in the analyses (Benros et al., 2016). In that case, it is important to figure out how badly these problems happen in empirical settings or where the problems might be more severe.

In this paper, we use causal graph theory to assess and discuss the fundamental methodological features (e.g., confounding, interaction effects between gene-gene, gene-environment, and gene-exposure) for causal association studies where the PRS is used as a covariate in the model to control for genetic background. Guidelines are provided to help researchers determine when and how it is relevant to include PRS as part of the causal statistical model. We will show the impact of wrongly including the PRS through an extensive series of simulation studies.

The manuscript is organized as follows. First, we will explain the basic concept of the PRS and discuss the use of directed acyclic graphs (DAGs) for illustrating the different scenarios. In section three, we will briefly overview the simulation settings, including analysis models. After the results section, we discussed the several scenarios including some empirical examples in which the association between exposure and outcome was biased when including the PRS in the association model, and then provide several recommendations for proper use of the PRS as covariate in the causal model.

2. Methods

2.1. Polygenic risk scores (PRS)

The PRS generally relies on genome-wide single-nucleotide polymorphism (SNP) data, in which a set of SNPs identified in a discovery GWAS (typically from a separate training sample) is used to construct the PRS (Agerbo et al., 2015; Dudbridge, 2013, 2016; Euesden et al., 2015; Maher, 2015). This polygenic score is then subsequently used as a covariate in a statistical model of interest to test or adjust for association testing or risk prediction (Agerbo et al., 2015; Dudbridge, 2013, 2016; Euesden et al., 2015; Maher, 2015).

Let us consider a sample of N individuals where $G_1, G_2, \dots, G_P$ be vectors of length N that contain the registration from P SNPs that were previously found to be associated with an outcome Y . Each of the P predictors has a corresponding estimated regression coefficient, $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_P$ (linear regression for the continuous trait or logistic regression for the binary trait) obtained from an external source and these are typically combined into a PRS using the formula (Dudbridge, 2013, 2016; Maher, 2015; Wray et al., 2014)

$$PRS = \sum [\hat{\beta}_{i, discovery} * G_{ij}],$$

or a variation thereof. Typically, the P vectors $G_1, G_2, \dots, G_P$ contain elements with the values 0, 1, or 2 or estimates of the number of disease alleles, and the selection procedure commonly prunes the number of associated SNPs such that they are uncorrelated or in linkage disequilibrium. While there are different approaches for creating the PRS, the exact definition and the relationship between the SNPs used to create it is not critical to our exposition. Without loss of generality, we can assume that the SNPs are independent and combined additively into the final score. A detailed description of the PRS can be found elsewhere (Dudbridge, 2013, 2016; Maher, 2015).

2.2. Theoretical based on causal graphs

Throughout the paper, we will use the following notations: X denotes the exposure (treatment) of interest, Y denotes the outcome (disease), C and U denote the measured/observed and unmeasured/unobserved confounding, respectively, and G denotes genotype. In the following paragraphs, we will first describe the fundamental notion of the directed acyclic graph (DAG), and then we will discuss the concept of confounding variables and the instrumental variable.

A DAG in which the edges are ordered pairs and the edges are arrows and shows the flow from one node (e.g., exposure) to another (e.g., outcome) and not vice versa, and it can be used to describe a causal diagram. For example, if X has a causal effect on Y , then X is a predecessor of the node Y (Fig. 1a). For a detailed explanation of the DAGs, we

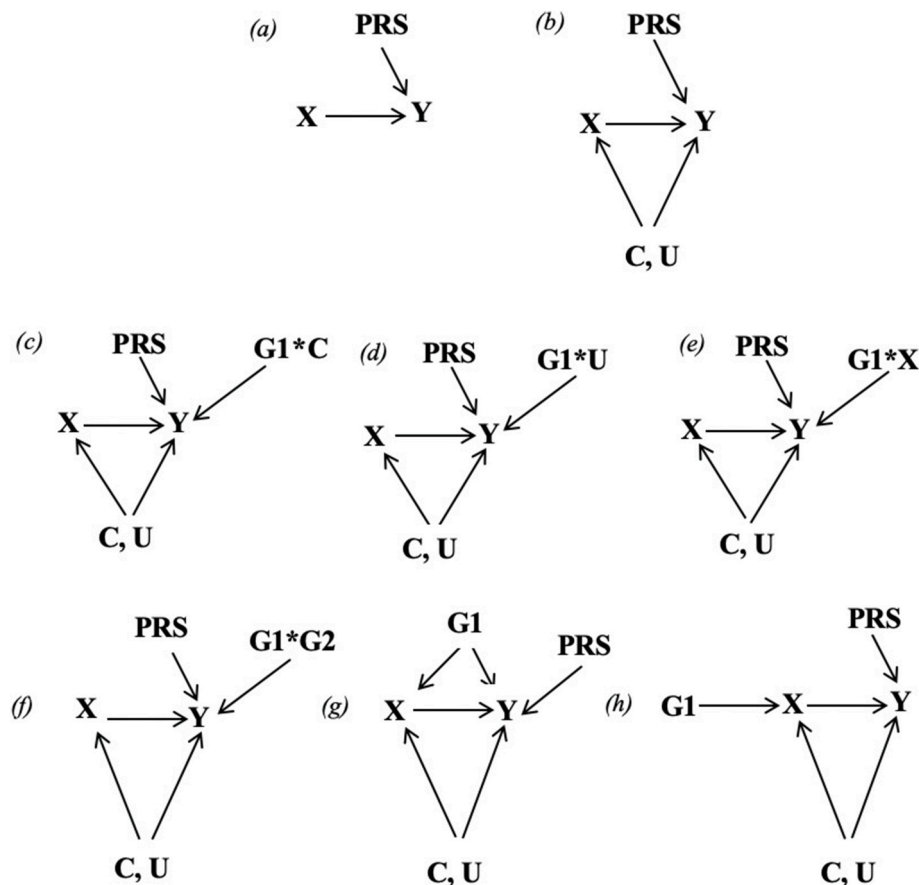


Fig. 2. Overview of the different scenarios of the simulations. Directed acyclic graphs (2a-h), where PRS (polygenic risk scores), X = exposure, Y = outcome, C = observed confounder and U = unobserved confounder, G = genotype.

refer to the literature, e.g., (Imbens, 2020; PEARL, 1995; VanderWeele and Robins, 2007). Moreover, VanderWeele et al. also provided findings that allow researchers to define the sign of bias when controlling for confounding is insufficient in specific cases (VanderWeele et al., 2008). Here, it suffices those causal relations between variables are represented by directed arrows from cause to effect and all causal relations of the exposure and outcome are represented.

A DAG is used to represent relationships among variables related to a causal hypothesis and it represents the researchers' idea about the mechanisms that generated the data (Sauer et al., 2013). DAGs are often used to determine acceptable adjustment procedures for epidemiological studies. However, building those is not straightforward. Previous studies have established the systematic method to develop DAGs (Ferguson et al., 2020) and confounder selection's principles (VanderWeele, 2019; VanderWeele and Shpitser, 2011). Ferguson et al. (2020) developed a unique theory-driven strategy for creating DAGs using empirical literature. Their strategy is divided into three phases: (i) each study's conclusions are 'mapped' into a DAG, (ii) the causal mechanisms in those DAGs are systematically analysed and rectified utilizing numerous causal inference concepts, (iii) the generated DAGs are subsequently combined to form one or more integrated DAGs.

A confounding variable (C or U) is often defined as a common cause of both the exposure (X) and the outcome (Y) (Fig. 1b). Here, X and Y share a common cause C or U. Moreover, VanderWeele and Shpitser defined confounder as "a pre-exposure covariate C for which there exists a set of other covariates Z such that effect of the exposure on the outcome is unconfounded conditional on (Z, C) but such that for no proper subset of (Z, C) is the effect of the exposure on the outcome unconfounded given the subset" (VanderWeele and Shpitser, 2013). On

the other hand, an instrumental variable is associated to the exposure variable and is independent of confounding variables. Despite the fact that it influences the exposure, which influences the outcome variable only through the exposure (Fig. 1c) (Greenland, 2000; Islam et al., 2022; Zohoori and Savitz, 1997). Here, G acts as an instrumental variable.

2.3. Simulation set-up

We used a simulation study to illustrate and assess the applicability and validity of the PRS or the genetic marker(s) (that are used for calculating PRS) as a covariate on the causal association model. We tried to follow the guidelines provided by Morris et al. (2019). In this section, we briefly describe the simulation settings, including different scenarios and data analyses.

In real life, researchers generally assume that the PRS is independent of other variables in the causal pathway. However, this assumption is not always true, and below, we explain several possible scenarios where using PRS as a covariate in the model may induce bias in the association between exposure and outcome.

Fig. 2 shows several DAGs that show realistic relationships that could occur in a GWAS setting and provides an overview of the different scenarios of the simulation settings. If a situation like Fig. 2a (like an RCT) in which there are no confounding variables and the PRS only affects the outcome or if there are confounders and we assume that the PRS completely independent of them (Fig. 2b), adjustment of the PRS in the model will not remove any bias in the association between exposure and outcome, but it may improve the efficiency of the estimates. However, if the PRS directly influences the confounders (Fig. 2c) or there is an interaction between a genotype and confounders (Fig. 2c and d),

Table 1

Overview of the simulation settings.

Settings	Scenarios
Sample size (n)	100, 500, 5000, 100000
Number of replications (r)	1000
Exposure (X)	Continuous
Outcome (Y)	Continuous
Confounding variables (C & U)	Continuous
Baseline effects (intercept) (α)	0.5 (arbitrary)
True exposure effect on outcome	1 (arbitrary)
Effect of measured confounder on the exposure and outcome (β_c)	0.2–1.0
Effect of unmeasured confounder on the exposure and outcome (β_u)	0.1–0.8
Correlation between C & U	0.40
Prevalence of U	10–25%
Effect of PRS on the outcome (β_{PRS})	Lognormal distribution with mean '0' and standard deviation '1'
Interaction effect of exposure and genotype on the outcome (β_{X*G_1})	1 or 2
Genotype	$G_1, G_2, \dots, G_{10}$
Frequency of the dominant allele in the population (p)	Uniform distribution, min = 0.05 & max = 0.95
Frequency of the recessive allele in the population (q)	1-p
Hardy-Weinberg (HW) equation	$1 = p^2 + 2pq + q^2$ p: frequency of homozygous dominant genotype (A) q: frequency of homozygous recessive genotype (a) 2pq: frequency of heterozygous genotype p + q = 1
Effect of genetic markers on the outcome	Lognormal distribution with mean '0' and standard deviation '1'
PRS calculation	$i) PRS = \sum_{i=1}^{10} G_i$ ii) $PRS = \sum_{j=1}^{10} \hat{\beta}_j * G_{ij}$

Table 2

The impact of different magnitudes of confounding effects on exposure effects.

Model	Estimate (SE) (True effect = 1.0)	Confidence Interval (LCL; UCL)	RMSE
Small effect of unmeasured confounding: $\beta_c = 0.2$ and $\beta_u = 0.1$			
Crude Model	1.04 (0.01)	1.02; 1.07	0.05
Model with PRS	1.05(0.01)	1.03; 1.06	0.05
Model with PRS and measured confounders	1.00 (0.01)	0.98; 1.02	0.01
Full Model	1.00 (0.00)	0.99; 1.01	0.00
Moderate effect of unmeasured confounding: $\beta_c = 0.5$ and $\beta_u = 0.25$			
Crude Model	1.23 (0.01)	1.20; 1.25	0.23
Model with PRS	1.23 (0.01)	1.22; 1.24	0.23
Model with PRS and measured confounders	1.01 (0.01)	1.00; 1.02	0.01
Full Model	1.00 (0.00)	0.99; 1.01	0.00
Strong effect of unmeasured confounding: $\beta_c = 0.8$ and $\beta_u = 0.4$			
Crude Model	1.90 (0.00)	1.90; 1.91	0.90
Model with PRS	1.90 (0.00)	1.90; 1.90	0.90
Model with PRS and measured confounders	1.73 (0.00)	1.73; 1.74	0.73
Full Model	1.00 (0.00)	0.99; 1.01	0.00

SE: Standard Error.

adjustment of PRS may induce high bias. Furthermore, if a genotype interacts with the exposure or another genotype (Fig. 2e and f), adjustment of PRS may induce high bias unless the interaction is considered in the model. In a study where a genotype acts as a confounding or instrumental variable (Fig. 2g and h), adjustment of the PRS in the model may amplify the bias.

We considered several realistic scenarios, as explained above, such as gene-environment (confounding variables), gene-gene, and gene-exposure interaction; when a gene acts as a confounder, the gene acts as an instrumental variable and replicated each scenario 1000 times with a variable number of $N = 100$ to 100000 observations. The

scenarios are illustrated in different DAGs in Fig. 2.

Scenario I: Fig. 2b represents scenario I, in which the PRS only affects the outcome Y. The confounding variables (C and U) are independent of the PRS. Moreover, there is no interaction between PRS and exposure.

Scenario II: Scenario II is like the scenario I, except there is an interaction between a genotype (e.g., G_1 , that is used for calculating the PRS) and the exposure (Fig. 2e)

Scenario III: In this scenario (Fig. 2c), there is an interaction between a genotype (e.g., G_1) and the measured confounding variables (C).

Scenario IV: Scenario IV (Fig. 2d) is like scenario III, but the interaction term is between a genotype (e.g., G_1) and the unmeasured confounding variables (U).

Scenario V: In this scenario (Fig. 2f), we examined the interaction between genotypes (e.g., both G_1 and G_2 are used for calculating the PRS).

Scenario VI: In this scenario (Fig. 2g), the genotypes (e.g., G_1) act as a measured confounding variable like C.

Scenario VII: We examined a setting that reflects a Mendelian randomization (MR) study in which the PRS can be adjusted. Here, the genotypes, e.g., G_1 , act as an instrumental variable (Fig. 2h).

We simulated continuous Y, X, a measured confounding C, and an unmeasured confounding U. The effect of measured confounder (β_c) and unmeasured confounder (β_u) on the exposure and outcome were simulated in the interval 0.2 to 1.0 and 0.1 to 0.8, respectively. We assumed that the C and U are correlated with a correlation coefficient of 0.40 and follow a multivariate normal distribution with zero and a variance of 1. Then we dichotomized the U with a prevalence of 10–25%. The allele frequencies for the Gs were simulated using the uniform distribution on the interval of 0.05–0.95 and subsequently used to produce genotypes (which by construction fulfill the Hardy-Weinberg principle), and the effects of the Gs were generated using the lognormal distribution with mean zero and standard deviation one. Moreover, we constructed the PRS using a sum of the allelic count or using a weighted sum of allelic count from the same data where the weights are given by regression coefficients from a univariate regression model (i.e., linear regression).

For continuous variables, the following linear models was used for generating data. The models are described in the following text:

Exposure model : $X = \alpha_x + \beta_c C + \beta_u U + \varepsilon_x$; where $\varepsilon_x \sim N(0, 1)$

Outcome model : $Y = \alpha_y + \beta_x X + \beta_c C + \beta_u U + \sum_{i=1}^{10} \beta_{g_i} G_i + \varepsilon_y$; where $\varepsilon_y \sim N(0, 1)$

Outcome model with interaction : $Y = \alpha_y + \beta_x X + \beta_c C + \beta_u U + \sum_{i=1}^{10} \beta_{g_i} G_i + \beta_{X*G_1} X * G_1 + \varepsilon_y$; where $\varepsilon_y \sim N(0, 1)$

Here, the item $\sum_{i=1}^{10} \beta_{g_i} G_i$ is the PRS and that the parameters here are given from the linear regression estimates.

Since, the data generation mechanism based on the linear model, the data analyses models are also considered as linear:

Crude model: $Y = \alpha_y + \beta_x X$

Model adjusted for PRS: $Y = \alpha_y + \beta_x X + \beta_{PRS} PRS$

Model adjusted for PRS and measured confounding: $Y = \alpha_y + \beta_x X + \beta_{PRS} PRS + \beta_c C$

Model adjusted for interaction between exposure and one of the genotypes, e.g., G_1 :

$Y = \alpha_y + \beta_x X + \beta_{PRS} PRS + \beta_c C + \beta_{X*G_1} X * G_1$

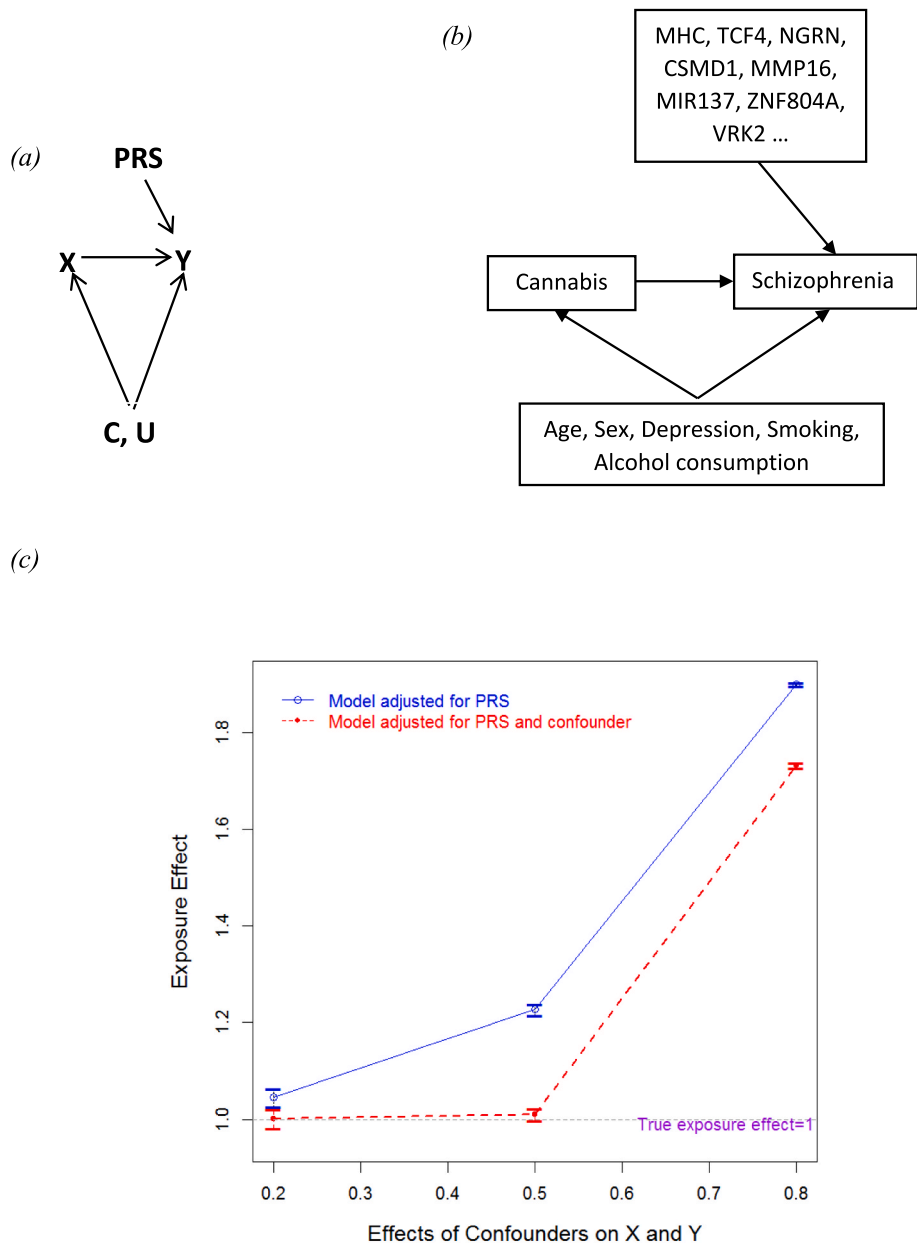


Fig. 3. Directed acyclic graphs (2b, where PRS = polygenic risk scores, X = exposure, Y = outcome, C = observed confounder and U = unobserved confounder) and plot of different effects of confounders vs. exposure effect. Here genotypes that were used for calculating the PRS were not associated with other variables in the causal pathway. The different points on the lines indicate different confounding effects on the exposure and outcome.

Table 3 Exposure effects in relation to gene-gene interaction.			
Model	Estimate (SE) (True effect = 1.0)	Confidence Interval (LCL; UCL)	RMSE
Crude Model	1.05 (0.02)	1.00; 1.09	0.05
Model with PRS	1.05 (0.02)	1.01; 1.08	0.05
Model with PRS and measured confounders	1.00 (0.02)	0.97; 1.04	0.02
Full Model	1.00 (0.01)	0.99; 1.01	0.01

Here the β_x represents the effect of exposure on the Y and β_c , β_u represent the effect of measured and unmeasured confounding, respectively on the X and Y. Moreover, α_x and α_y represent the intercept of the X and Y, respectively. β_{prs} and $\beta_{x \times G1}$ represent the effect of PRS and interaction effect of exposure and genotype on the outcome. It should be

Table 4 Exposure effects in relation to genotype exposure interaction.			
Model	Estimate (SE) (True effect = 1.0)	Confidence Interval (LCL; UCL)	RMSE
Crude Model	1.65 (0.01)	1.63; 1.67	0.66
Model with PRS	1.65 (0.01)	1.64; 1.67	0.66
Model with PRS and measured confounders	1.56 (0.01)	1.54; 1.58	0.56
Model with PRS, measured confounders and interaction	1.01 (0.01)	0.99; 1.02	0.01
Full Model (without interaction)	1.55 (0.01)	1.54; 1.56	0.55

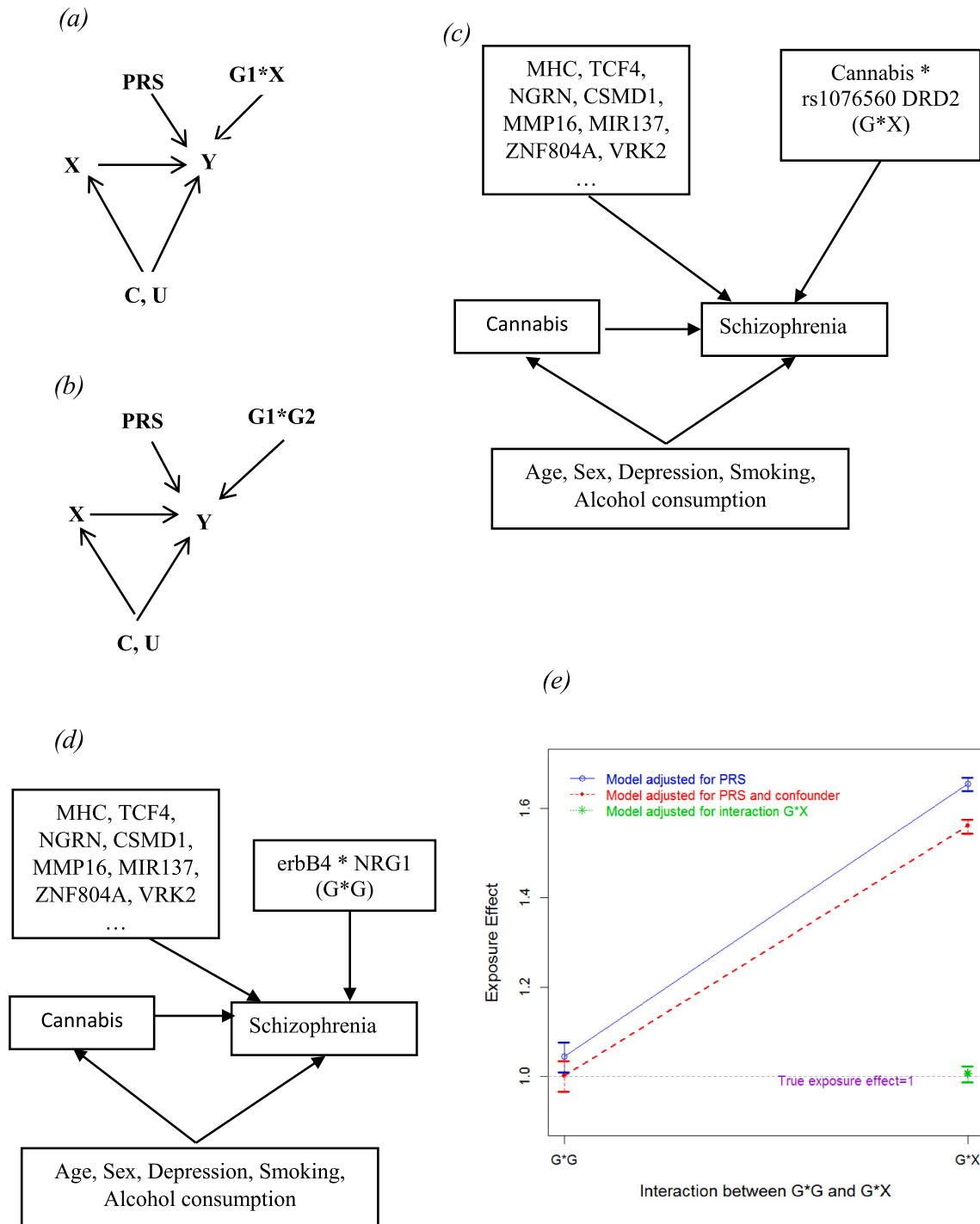


Fig. 4. Directed acyclic graphs (2e and 2f, where PRS = polygenic risk scores, G1 and G2 = two of the genotypes used for calculating the PRS, X = exposure, Y = outcome, C = observed confounder and U = unobserved confounder) and plot of the effects of gene*gene and gene*exposure interaction vs. exposure effect. Here genotypes that were used for calculating the PRS were not associated with other variables in the causal pathway.

noted that the simulated data were analysed using the linear regression. In all analyses, the confounding U was not included in the model since it was assumed unmeasured. The crude and adjusted exposure effects were estimated in each simulated dataset and averaged across simulations. In the crude model, only the exposure X and in the adjusted models, C and PRS were considered. The bias of exposure effects was defined as the difference between the mean of exposure effects based on 1000 simulation runs, and the true exposure effect (e.g., $\beta = 1$). The confidence interval (CI) was estimated by 2.5 and 97.5 percentiles of the 1000 estimates. We also estimated the root-mean-square error (RMSE) of

estimates. We used statistical software R (Windows, version 3.4.2) to simulate and analyse the data. The summary of the simulation settings including different scenarios are illustrated in Table 1 and Fig. 2. We have added the codes used for the simulation (Supplementary file 2).

3. Results

Table 2 presents the simulation results based on the scenario I (Fig. 2b) where we assessed the impact of different magnitudes of confounding effects. It showed that the bias (the estimates of exposure were

Table 5

Exposure effects in relation to genotype used for calculating the PRS interacts with the measured confounding variable (C).

Model	Estimate (SE) (True effect = 1.0)	Confidence Interval (LCL; UCL)	RMSE
Crude Model	1.94 (0.02)	1.90; 1.97	0.94
Model with PRS	1.94 (0.02)	1.91; 1.97	0.94
Model with PRS and measured confounders	1.00 (0.01)	0.99; 1.02	0.01
Full Model	1.00 (0.01)	0.99; 1.01	0.01

Table 6

Exposure effects in relation to genotype used for calculating the PRS interacts with the unmeasured confounding variables (U).

Model	Estimate (SE) (True effect = 1.0)	Confidence Interval (LCL; UCL)	RMSE
Interaction effect: $\beta = 2$			
Crude Model	1.17 (0.01)	1.15; 1.18	0.17
Model with PRS	1.17 (0.01)	1.16; 1.18	0.17
Model with PRS and measured confounders	1.05 (0.01)	1.04; 1.06	0.05
Full Model	1.00 (0.01)	0.99; 1.01	0.00
Interaction effect: $\beta = 4$			
Crude Model	1.34 (0.01)	1.31; 1.37	0.34
Model with PRS	1.34 (0.01)	1.32; 1.36	0.34
Model with PRS and measured confounders	1.12 (0.01)	1.10; 1.15	0.12
Full Model	1.00 (0.00)	0.99; 1.01	0.00

over-estimated and far from the true value) increased from 1.00 [0.98–1.02], RMSE = 0.01 to 1.73 [1.73–1.74], RMSE = 0.73, with increasing the strength, $\beta_c = 0.2$ to 0.8 and $\beta_u = 0.1$ to 0.4, of the confounding variables (Fig. 3c). This phenomenon is not new for such studies, this is generally happened for all observational studies.

Moreover, we found that the estimates are similar for both crude and the model with PRS, and for strong unmeasured confounding, a significant amount of bias is present even adjustment of the measured confounders. Also, as expected, the variation of the estimates was larger for the smallest sample size (Supplementary Table S1).

Tables 3 and 4 presents gene-gene interaction (Fig. 2f) or genotype exposure interaction (Fig. 2e), respectively. We observed that the amount of bias was high when the genotype interacts with the exposure, 1.56 [1.54–1.58], RMSE = 0.56. But as expected, when we considered the interaction effect in the analyses, the estimates were similar with the true value, 1.01 [0.99–1.02], RMSE = 0.01. However, without adjustment of gene-gene interaction, the amount of bias almost zero (Fig. 4e). This is due to the additive and/or weak interaction effects in our simulated data.

When a genotype (e.g., G_1) that has been used for calculating the PRS interacts with the measured confound variable (C) (Fig. 2c), the exposure effect similar with the true value ($\beta_x = 1$) if the interaction effect was considering in the analysis (Table 5).

However, a significant amount of bias was present even after adjustment for the measured confounding effects when the genotype interacts with the unmeasured confounding variables (U) (Figs. 2d), 1.05 [1.04–1.06], RMSE = 0.05, and this bias amplified, 1.12 [1.10–1.15], RMSE = 0.12, when the interaction effect was two times stronger (Table 6 and Fig. 5e). In that case, the bias could be more if the unmeasured confounding was independent of the measured confounders (data not shown).

Moreover, Tables 7 and 8 represents the sceneries when the genotype (e.g., G_1) acts as a measured confounder (Fig. 2g) or an instrumental variable (Fig. 2f), respectively. We found the exposure effects were biased when the genotype acts as a measured confounder, 0.59 [0.57–0.60], RMSE = 0.415 or an instrumental variable, 0.69 [0.68–0.71], (Fig. 6e). In that case, the bias disappeared when we

condition the G_1 on the model (data not shown).

4. Discussion

Our simulation study shows that the PRS can be used as a covariate in the causal association model to adjust for genetic background or to increase the precision or to reduce the systematic bias of the exposure effect in psychological studies. However, we found in some scenarios, such as when a genotype that has been used for calculating the PRS interact with the exposure or confounding factors; acts as a confounding variable or an instrumental variable, the exposure effects were biased. In all situations except the scenarios illustrated in Fig. 6, the exposure effects are over-estimated and consequently, the direction of the bias was positive. In that case, the amount of bias was much higher, and direction of bias was negative when the genotype acts as a confounding. Moreover, as expected the magnitude of bias was less or disappeared when the interactions effects (e.g., gene-exposure interaction or gene-measured confounding interaction) were considered in the model.

When a genotype interacts with the exposure of interest and used for calculating the PRS for the outcome (e.g., PRS for schizophrenia), the exposure effect was biased unless the interaction effect was considering in the causal model (Fig. 4e). For example, in real life, several studies (Costas et al., 2011; Henquet et al., 2009; Nieman et al., 2016; Vaessen et al., 2018) showed that the catechol O-methyltransferase (COMT) (Zammit et al., 2007) genotype or dopamine D2 receptor (DRD2) genotype (Colizzi et al., 2015) interact with the cannabis use. Moreover, COMT is also a candidate gene for the schizophrenia (Craddock et al., 2006; Daniel Weinberger, n.d.; Glatt et al., 2003). In that case, if these genotypes (i.e., COMT) is used for calculating the PRS for schizophrenia and the PRS used as a covariate in the cannabis-schizophrenia association model, without account for the COMT*cannabis interaction, the effect of cannabis on the schizophrenia will be biased.

Gene and Gene interaction did not induce any bias on the exposure effect in our simulation (Fig. 4e). This may be because we used linear model for simulation and analysis of the data. In such case, a real example, Xu et al. (2007) showed that DRD2 gene interact with the COMT gene and both genes also candidate gene for the schizophrenia (Craddock et al., 2006; Daniel Weinberger, n.d.; Glatt et al., 2003; Hoenicka et al., 2006). Hence, if someone calculate the PRS for schizophrenia where these two genes were used, as we showed in Fig. 4e, without adjusting for this DRD2-COMT gene interaction, the cannabis-schizophrenia association may not be biased given that the association between gene and outcome is linear.

As explained above, COMT or DRD2 associated with the cannabis use as well as the risk of schizophrenia. So, these genes may play a role like confounding variables in the cannabis-schizophrenia model. We found that when a genotype acts as a confounder and simultaneously is used as part of the PRS, the estimates were biased (Fig. 6e) and underestimated. Therefore, theoretical knowledge and previous research findings should be used to justify those genotypes used for calculating the PRS are not acting as confounders though it is a cumbersome work to identify as thousands of genotypes used for the calculation of the PRS.

We also found that when the PRS is only used as a covariate in the model and assumed that the genotypes that were used for calculating the PRS were not associated with other variables in the causal pathway (Fig. 2a), the exposure effect is unbiased given that the effect of confounding is not strong. However, in the case of strong confounding effects or interaction between genotype and confounders (measured or unmeasured), the estimates become biased except for the scenario where the interaction between measured confounding and genotype was considering in the model. Again, when studying the effect of cannabis use and the risk of schizophrenia, DRD2 gene interacts with the early life stress (a confounding for the cannabis-schizophrenia association) (Klaus et al., 2017). Hence, without accounting for the interaction between DRD2 and life stress, the cannabis effect may be biased.

If a genotype purely acts as an instrumental variable (a variable

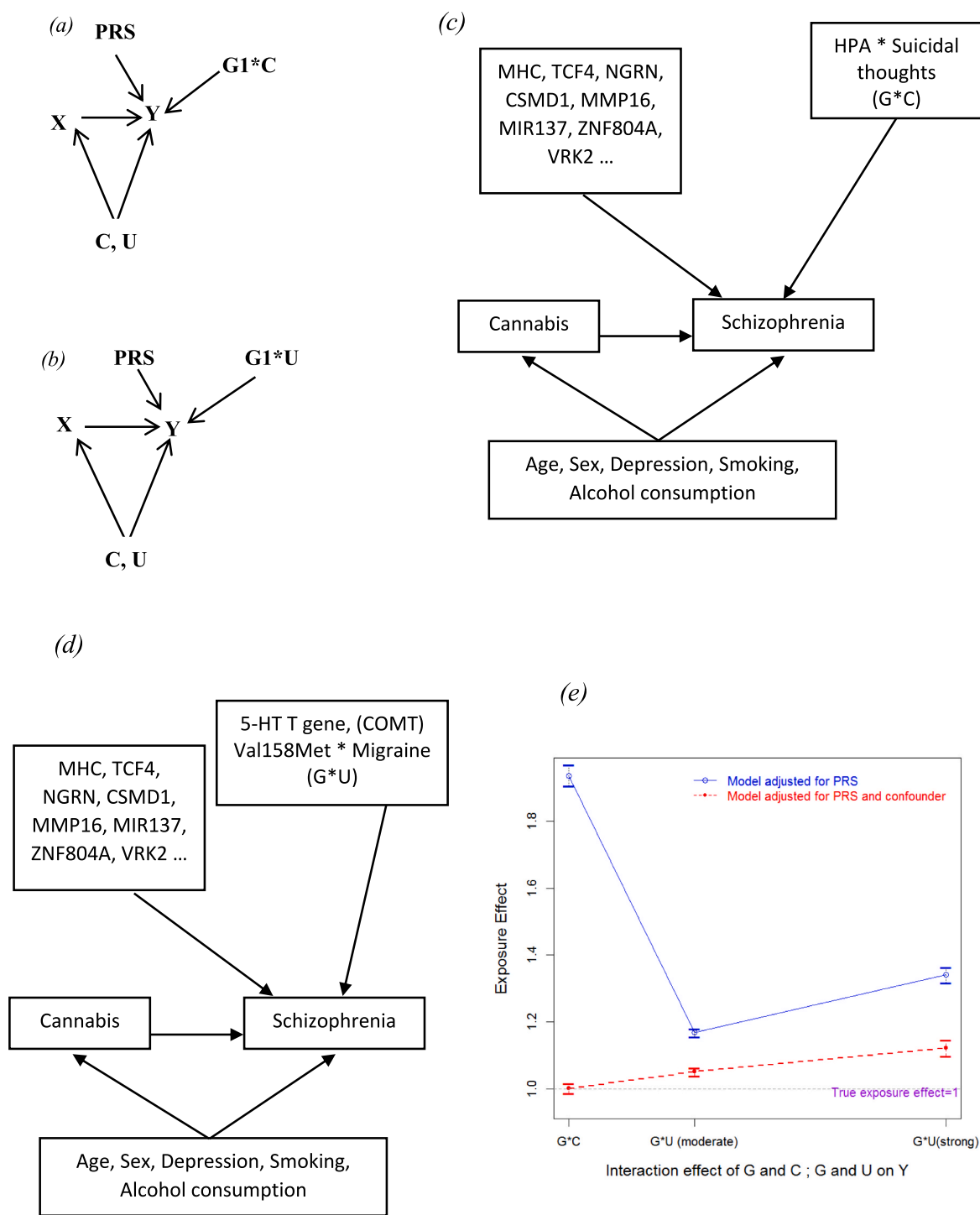


Fig. 5. Directed acyclic graphs (2c and 2d, where PRS = polygenic risk scores, G1 = one of the genotypes used for calculating the PRS, X = exposure, Y = outcome, C = observed confounder and U = unobserved confounder) and plot of the effects of gene*measured confounding and gene*unmeasured confounding (moderate and strong) interaction vs. exposure effect. Here genotypes that were used for calculating the PRS were not associated with other variables in the causal pathway.

Table 7
Exposure effects in relation to genotype acts as a measured confounder.

Model	Estimate (SE) (True effect = 1.0)	Confidence Interval (LCL; UCL)	RMSE
Crude Model	1.18 (0.01)	1.16; 1.20	0.18
Model with PRS	0.63 (0.01)	0.61; 0.64	0.37
Model with PRS and measured confounders	0.59 (0.01)	0.57; 0.60	0.41
Full Model	1.00 (0.00)	0.99; 1.01	0.00

Table 8
Exposure effects in relation to genotype acts as an instrumental variable.

Model	Estimate (SE) (True effect = 1.0)	Confidence Interval (LCL; UCL)	RMSE
Crude Model	1.05 (0.02)	1.00; 1.09	0.05
Model with PRS	1.05 (0.02)	1.01; 1.08	0.05
Model with PRS and measured confounders	1.00 (0.02)	0.97; 1.04	0.02
Full Model	1.00 (0.00)	0.99; 1.01	0.00

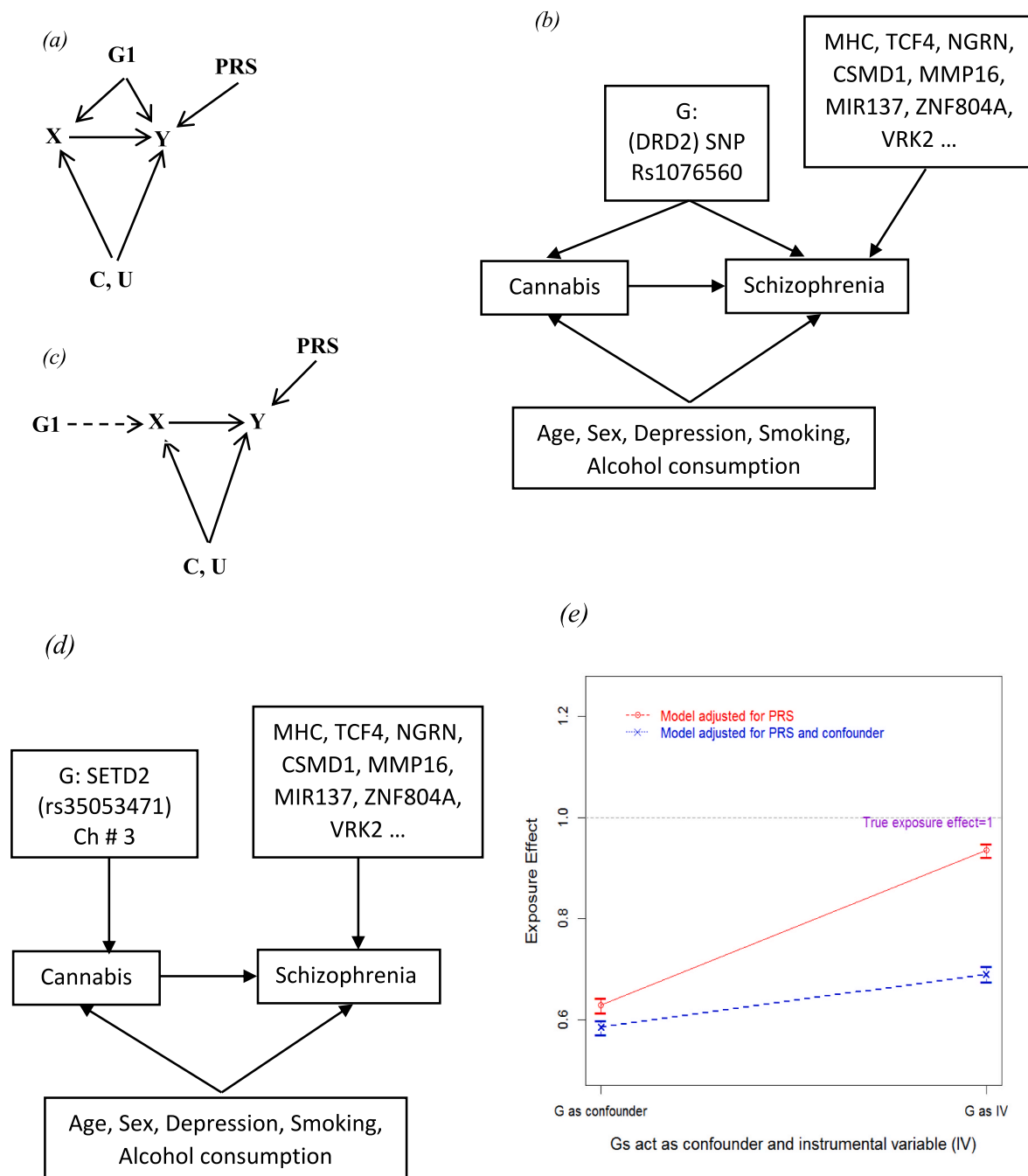


Fig. 6. Directed acyclic graphs (2g and 2h, where PRS = polygenic risk scores, G1 = one of the genotypes used for calculating the PRS and acts as a confounders and instrumental variables, X = exposure, Y = outcome, C = observed confounder and U = unobserved confounder) and plot of the effects of gene acts as confounding and instrumental variable.

associated with the exposure and affect the outcome only through the exposure) but used for calculating the PRS, the exposure effects far from the true value and underestimated (Fig. 6e). In that case, a classical regression method is not recommended, and a Mendelian randomization (MR) study should be considered.

Our study has several strengths: first, we conducted an extensive simulation study using a wide range of realistic scenarios that reflect in real-life example; second, we described each scenario in the DAGs that is helpful for understanding the real-life setting. A limitation of our simulation study is that we restricted the simulations to a continuous outcome and continuous PRS with linear model. We chose this approach, because we assumed that the findings for binary outcome may in line with the continuous outcome. Future research could extend the

simulations to settings with binary outcome. Another limitation of our study is we did not provide any application based on real data because of our resource limitation.

Guidelines for proper application of PRS.

- Researchers should carefully decide whether adjustment for genetic background is an essential and clinically relevant for their causal hypothesis because the PRS depends heavily on specific disease (Torkamani et al., 2018) and unnecessary adjustment may induce bias on the exposure effect.
- Both DAG and expert knowledge should be used to understand the causal mechanism between exposure, outcome, confounding, and genotypes and proper statistical models. When sufficient knowledge

is available to construct a DAG, it helps the investigator to better understand statistical associations observed in the data.

- It is important to identify genotypes that act as confounding or interact with the exposure or measured confounding variables though this a cumbersome work as there are thousands of SNPs used for calculating the PRS. If it is possible to identify a genotype that acts as a confounding or instrumental variable, that should not be used for calculating the PRS for the outcome.
- If the exposure effect changed significantly after adjusting for the PRS, researchers may think the possible issues (e.g., Gene acts as confounding, gene-exposure, or gene-confounding interaction) and statistical model fitting criteria in their analyses.

5. Conclusions

Our study shows that caution should be taken when the PRS is used as a covariate to adjust for genetic background or to increase the precision or reduce the systematic bias of the exposure effect in psychological studies. Theoretical knowledge should be applied when deciding upon the statistical model as the PRS may induces bias in the causal association between exposure and outcome in some scenarios. We argue that our study will help to improve the understanding and validity of the assumptions underlying the causal association between environmental, genetic factors and disease, and drug response, genetic factors, and disease.

Ethics approval and consent to participate

Not applicable.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' contributions

MJU reviewed the literature, analysed the data, and drafted the first copy of the manuscript. TA contributed significantly writing the manuscript. CH, MN, and CTE assisted in the writing of the manuscript and developed it. CTE has provided instructive assistance throughout the process and designed the study, made substantial contributions to the interpretation, and oversaw the study project. The manuscript has been read and accepted by all the contributors.

Availability of data and materials

The data, that support the findings of this study, was simulated and there was no real human genetic data. Moreover, the simulation process has been described in the "Methods" section. No additional supporting data are required for this study.

Consent for publication

Not applicable.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Not applicable.

List of abbreviations

PRS	Polygenic Risk Score
GWAS	Genome-Wide Association Studies
CI	Confidence Interval
DAG	Directed Acyclic Graph
SE	Standard Error
SNP	Single-Nucleotide Polymorphism
RMSE	Root-Mean-Square Error
COMT	Catechol O-methyltransferase

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.metip.2022.100099>.

References

- Agerbo, E., Byrne, M., Eaton, W.W., Mortensen, P.B., 2004. Marital and labor market status in the long run in schizophrenia. *Arch. Gen. Psychiatr.* 61 (1), 28–33.
- Agerbo, E., Sullivan, P.F., Vilhjálmsson, B.J., Pedersen, C.B., Mors, O., Børglum, A.D., Hougaard, D.M., Hollegaard, M.V., Meier, S., Mattheisen, M., Ripke, S., Wray, N.R., Mortensen, P.B., Vilhjálmsson, B.J., Pedersen, C.B., Mors, O., Børglum, A.D., Hougaard, D.M., Hollegaard, M.V., et al., 2015. Polygenic risk score, parental socioeconomic status, family history of psychiatric disorders, and the risk for schizophrenia. *JAMA Psychiatr.* 72 (7), 635–641. <https://doi.org/10.1001/jamapsychiatry.2015.0346>.
- Anderson, J.S., Shade, J., DiBlasi, E., Shabalin, A.A., Docherty, A.R., 2019. Polygenic risk scoring and prediction of mental health outcomes. *Curr. Opin. Psychol.* 27, 77–81. <https://doi.org/10.1016/j.copsyc.2018.09.002>.
- Benca, C.E., Derringer, J.L., Corley, R.P., Young, S.E., Keller, M.C., Hewitt, J.K., Friedman, N.P., 2017. Predicting cognitive executive functioning with polygenic risk scores for psychiatric disorders. *Behav. Genet.* 47 (1), 11–24. <https://doi.org/10.1007/s10519-016-9814-2>.
- Benros, M.E., Trabjerg, B.B., Meier, S., Mattheisen, M., Mortensen, P.B., Mors, O., Børglum, A.D., Hougaard, D.M., Nørgaard-Pedersen, B., Nordentoft, M., Agerbo, E., 2016. Influence of polygenic risk scores on the association between infections and schizophrenia. *Biol. Psychiatr.* 80 (8), 609–616. <https://doi.org/10.1016/j.biopsych.2016.04.008>.
- Bogdan, R., Baranger, D.A.A., Agrawal, A., 2018. Polygenic risk scores in clinical psychology: bridging genomic risk to individual differences. *Annu. Rev. Clin. Psychol.* 14 (1), 119–157. <https://doi.org/10.1146/annurev-clinpsy-050817-084847>.
- Cecile, A., Janssens, W., John, P., Ioannidis, A., Bedrosian, S., Boffetta, P., Dolan, S.M., Dowling, N., Fortier, I., Freedman, A.N., 2011. Strengthening the reporting of genetic risk prediction studies (GRIPS): explanation and elaboration. *Eur. J. Epidemiol.* 26 (4), 313.
- Chun, S., Imakaev, M., Hui, D., Patsopoulos, N.A., Neale, B.M., Kathiresan, S., Stitzel, N. O., Sunyaev, S.R., 2018. Non-parametric polygenic risk prediction using partitioned GWAS summary statistics. In: *bioRxiv*. Cold Spring Harbor Laboratory, 370064. <https://doi.org/10.1101/370064>.
- Clare, K.E., Miller, M.H., Dillon, J.F., 2017. Genetic factors influencing drug-induced liver injury: do they have a role in prevention and diagnosis? *Curr. Hepatol. Rep.* 16 (3), 258–264.
- Colizzi, M., Iyegbe, C., Powell, J., Ursini, G., Porcelli, A., Bonvino, A., Taurisano, P., Romano, R., Masellis, R., Blasi, G., 2015. Interaction between Functional Genetic Variation of DRD2 and Cannabis Use on Risk of Psychosis. *Schizophrenia Bulletin*. Oxford University Press US.
- Compton, M.T., Manseau, M., 2017. The Complex Connection between Cannabis and Schizophrenia. Academic Press. <https://doi.org/10.1016/C2015-0-02279-1>.
- Costas, J., Sanjuán, J., Ramos-Ríos, R., Paz, E., Agra, S., Tolosa, A., Páramo, M., Brenlla, J., Arrojo, M., 2011. Interaction between COMT haplotypes and cannabis in schizophrenia: a case-only study in two samples from Spain. *Schizophr. Res.* 127 (1), 22–27.
- Craddock, N., Owen, M.J., O'donovan, M.C., 2006. The catechol-O-methyl transferase (COMT) gene as a candidate for psychiatric phenotypes: evidence and lessons. *Mol. Psychiatr.* 11 (5), 446.
- Daniel Weinberger. (n.d.). COMT and Schizophrenia.
- Dima, D., Breen, G., 2015. Polygenic risk scores in imaging genetics: usefulness and applications. *J. Psychopharmacol.* 29 (8), 867–871. <https://doi.org/10.1177/0269881115584470>.
- Dudbridge, F., 2013. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* 9 (3), e1003348.
- Dudbridge, F., 2016. Polygenic epidemiology. *Genet. Epidemiol.* 40 (4), 268–272.
- Escott-Price, V., Shuai, M., Pither, R., Williams, J., Hardy, J., 2017. Polygenic score prediction captures nearly all common genetic risk for Alzheimer's disease. *Neurobiol. Aging* 49, 214. <https://doi.org/10.1016/j.neurobiolaging.2016.07.018>.
- Euesden, J., Lewis, C.M., O'Reilly, P.F., 2015. PRSice: polygenic risk score software. *Bioinformatics* 31 (9), 1466–1468. <https://doi.org/10.1093/bioinformatics/btu848>.

- Ferguson, K.D., McCann, M., Katikireddi, S.V., Thomson, H., Green, M.J., Smith, D.J., Lewsey, J.D., 2020. Evidence synthesis for constructing directed acyclic graphs (ESCDAGs): a novel and systematic method for building directed acyclic graphs. *Int. J. Epidemiol.* 49 (1) <https://doi.org/10.1093/ije/dyz150>.
- Gattaz, W.F., Abrahão, W.L., Foccacia, R., 2002. Environmental risk factors of psychosis. In: *Risk and Protective Factors in Schizophrenia*, pp. 133–138. https://doi.org/10.1007/978-3-642-57516-7_12. Steinkopf.
- Glatt, S.J., Faraone, S.V., Tsuang, M.T., 2003. Association between a functional catechol O-methyltransferase gene polymorphism and schizophrenia: meta-analysis of case-control and family-based studies. *Am. J. Psychiatr.* 160 (3), 469–476.
- Greenland, S., 2000. An introduction to instrumental variables for epidemiologists. *Int. J. Epidemiol.* 29 (4), 722–729. <https://doi.org/10.1093/ije/29.4.722>.
- Henquet, C., Rosa, A., Delespaul, P., Papiol, S., Fañanás, L., Van Os, J., Myin-Germeys, I., 2009. COMT Val158Met moderation of cannabis-induced psychosis: a momentary assessment study of 'switching on' hallucinations in the flow of daily life. *Acta Psychiatr. Scand.* 119 (2), 156–160.
- Hoenicka, J., Aragüés, M., Rodríguez-Jimenez, R., Ponce, G., Martínez, I., Rubio, G., Jiménez-Arriero, M.A., Palomo, T., Parg, P., A, R.G., 2006. C957T DRD2 polymorphism is associated with schizophrenia in Spanish patients. *Acta Psychiatr. Scand.* 114 (6), 435–438.
- Imbens, G.W., 2020. Potential outcome and directed acyclic graph approaches to causality: relevance for empirical practice in economics. *J. Econ. Lit.* 58 (4) <https://doi.org/10.1257/JEL.20191597>.
- Islam, S.N., Ahammed, T., Anjum, A., Albalawi, O., Uddin, Md J., 2022. Reporting methodological issues of the mendelian randomization studies in health and medical research: a systematic review. *BMC Med. Res. Methodol.* 22 (1), 21. <https://doi.org/10.1186/s12874-022-01504-0>.
- Klaus, K., Butler, K., Durrant, S.J., Ali, M., Inglehearn, C.F., Hodgson, T.L., Gutierrez, H., Pennington, K., 2017. The effect of COMT Val158Met and DRD2 C957T polymorphisms on executive function and the impact of early life stress. *Brain Behav.* 7 (5).
- Lally, J., MacCabe, J.H., 2016. Personalised approaches to pharmacotherapy for schizophrenia. *BJPsych Adv.* 22 (2), 78–86. <https://doi.org/10.1192/apt.bp.114.013433>.
- Lewis, C.M., Vassos, E., 2017. Prospects for using risk scores in polygenic medicine. *Genome Med.* 9 (1), 96. <https://doi.org/10.1186/s13073-017-0489-y>.
- Maher, B.S., 2015. Polygenic scores in epidemiology: risk prediction, etiology, and clinical utility. *Curr. Epidemiol. Rep.* 2 (4), 239–244.
- Manrique-Garcia, E., Zammit, S., Dalman, C., Hemmingsson, T., Andreasson, S., Allebeck, P., 2014. Prognosis of schizophrenia in persons with and without a history of cannabis use. *Psychol. Med.* 44 (12), 2513–2521.
- Marangoni, C., Faedda, G.L., Baldessarini, R.J., 2018. Clinical and environmental risk factors for bipolar disorder: review of prospective studies. *Harv. Rev. Psychiatr.* 26 (1), 1–7. <https://doi.org/10.1097/HRP.0000000000000161>.
- Mistry, S., Harrison, J.R., Smith, D.J., Escott-Price, V., Zammit, S., 2018a. The use of polygenic risk scores to identify phenotypes associated with genetic risk of bipolar disorder and depression: a systematic review. *J. Affect. Disord.* 234, 148–155. <https://doi.org/10.1016/j.jad.2018.02.005>.
- Mistry, S., Harrison, J.R., Smith, D.J., Escott-Price, V., Zammit, S., 2018b. The use of polygenic risk scores to identify phenotypes associated with genetic risk of schizophrenia: systematic review. *Schizophr. Res.* 197, 2–8. <https://doi.org/10.1016/j.schres.2017.10.037>.
- Morris, T.P., White, I.R., Crowther, M.J., 2019. Using simulation studies to evaluate statistical methods. *Stat. Med.* 38 (11) <https://doi.org/10.1002/sim.8086>.
- Nieman, D.H., Dragt, S., van Duin, E.D.A., Denneman, N., Overbeek, J.M., de Haan, L., Rietdijk, J., Ising, H.K., Klaassen, R.M.C., van Amelsvoort, T., 2016. COMT Val158Met genotype and cannabis use in people with an at Risk Mental State for psychosis: exploring Gene x Environment interactions. *Schizophr. Res.* 174 (1), 24–28.
- Ott, J., 2015. Polygenic models for risk prediction in human genetics. *Hum. Hered.* 80 (4), 162–164.
- Pearl, J., 1995. Causal diagrams for empirical research. *Biometrika* 82 (4), 669–688. <https://doi.org/10.1093/biomet/82.4.669>.
- Power, R.A., Steinberg, S., Bjornsdottir, G., Rietveld, C.A., Abdellaoui, A., Nivard, M.M., Johannesson, M., Galesloot, T.E., Hottenga, J.J., Willemsen, G., 2015. Polygenic risk scores for schizophrenia and bipolar disorder predict creativity. *Nat. Neurosci.* 18 (7), 953–955.
- Pries, L.K., Guloksuz, S., Ten Have, M., De Graaf, R., Van Dorsselaer, S., Gunther, N., Rauschenberg, C., Reininghaus, U., Radhakrishnan, R., Bak, M., Rutten, B.P.F., Van Os, J., 2018. Evidence that environmental and familial risks for psychosis additively impact a multidimensional subthreshold psychosis syndrome. *Schizophr. Bull.* 44 (4), 710–719. <https://doi.org/10.1093/schbul/sby051>.
- Richardson, L., Hameed, Y., Perez, J., Jones, P.B., Kirkbride, J.B., 2018. Association of environment with the risk of developing psychotic disorders in rural populations: findings from the social epidemiology of psychoses in east anglia study. *JAMA Psychiatr.* 75 (1), 75–83. <https://doi.org/10.1001/jamapsychiatry.2017.3582>.
- Ripke, S., Neale, B.M., Pers, T.H., Julià, A., Kahn, R.S., Kalaydjieva, L., Karachanak-Yankova, S., Karjalainen, J., Kavanagh, D., Keller, M.C., 2014. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511 (7510), 421.
- Sauer, B., Brookhart, M.A., Roy, J.A., Vanderweele, T.J., 2013. Covariate selection. In: *Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide*. Agency for Healthcare Research and Quality (US).
- So, H.-C., Sham, P.C., 2017. Improving polygenic risk prediction from summary statistics by an empirical Bayes approach. *Sci. Rep.* 7 (1), 41262 <https://doi.org/10.1038/srep41262>.
- Torkamani, A., Wineinger, N.E., Topol, E.J., 2018. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* 19 (9), 581–590. <https://doi.org/10.1038/s41576-018-0018-x>.
- Trotta, A., Iyegbe, C., Di Forti, M., Sham, P.C., Campbell, D.D., Cherny, S.S., Mondelli, V., Aitchison, K.J., Murray, R.M., Vassos, E., 2016. Interplay between schizophrenia polygenic risk score and childhood adversity in first-presentation psychotic disorder: a pilot study. *PLoS One* 11 (9), e0163319.
- Vaessen, T.S.J., de Jong, L., Schäfer, A.T., Damen, T., Uittenboogaard, A., Krolinski, P., Nwosu, C.V., Pinckaers, F.M.E., Rotee, I.L.M., Smeets, A.P.W., Ermiş, A., Kennedy, J. L., Nieman, D.H., Tiwari, A., van Os, J., Drukker, M., 2018. The interaction between cannabis use and the Val158Met polymorphism of the COMT gene in psychosis: a transdiagnostic meta-analysis. *PLoS One* 13 (2), e0192658. <https://doi.org/10.1371/journal.pone.0192658>.
- VanderWeele, T.J., 2019. Principles of confounder selection. *Eur. J. Epidemiol.* 34 (3) <https://doi.org/10.1007/s10654-019-00494-6>.
- VanderWeele, T.J., Hernán, M.A., Robins, J.M., 2008. Causal directed acyclic graphs and the direction of unmeasured confounding bias. *Epidemiology* 19 (5). <https://doi.org/10.1097/EDE.0b013e3181810e29>.
- VanderWeele, T.J., Robins, J.M., 2007. Directed acyclic graphs, sufficient causes, and the properties of conditioning on a common effect. *Am. J. Epidemiol.* 166 (9) <https://doi.org/10.1093/aje/kwm179>.
- VanderWeele, T.J., Shpitser, I., 2011. A new criterion for confounder selection. *Biometrics* 67 (4). <https://doi.org/10.1111/j.1541-0420.2011.01619.x>.
- VanderWeele, T.J., Shpitser, I., 2013. On the definition of a confounder. *Ann. Stat.* 41 (1) <https://doi.org/10.1214/12-AOS1058>.
- Vilain, J., Galliot, A.M., Durand-Roger, J., Leboyer, M., Llorca, P.M., Schürhoff, F., Szöke, A., 2013. Environmental risk factors for schizophrenia: a review. *Encephale* 39 (1), 19–28. <https://doi.org/10.1016/j.encep.2011.12.007>.
- Vilhjálmsdóttir, B.J., Yang, J., Finucane, H.K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., Do, R., 2015. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* 97 (4), 576–592.
- Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.L., Brown, M.A., Yang, J., 2017. 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* 101 (1), 5–22.
- Walter, S., Glymour, M.M., Koenen, K., Liang, L., Tchetgen, E.J.T., Cornelis, M., Chang, S.-C., Rimm, E., Kawachi, I., Kubzansky, L.D., 2013. Performance of polygenic scores for predicting phobic anxiety. *PLoS One* 8 (11), e80326.
- Wray, N.R., Lee, S.H., Mehta, D., Vinkhuyzen, A.A.E., Dudbridge, F., Middeldorp, C.M., 2014. Research Review: polygenic methods and their application to psychiatric traits. *JCPP (J. Child Psychol. Psychiatry)* 55 (10), 1068–1087. <https://doi.org/10.1111/jcpp.12295>.
- Xu, H., Kellendonk, C.B., Simpson, E.H., Keilp, J.G., Bruder, G.E., Polan, H.J., Kandel, E. R., Gilliam, T.C., 2007. DRD2 C957T polymorphism interacts with the COMT Val158Met polymorphism in human working memory ability. *Schizophr. Res.* 90 (1), 104–107.
- Zammit, S., Spurlock, G., Williams, H., Norton, N., Williams, N., O'Donovan, M.C., Owen, M.J., 2007. Genotype effects of CHRNA7, CNR1 and COMT in schizophrenia: interactions with tobacco and cannabis use. *Br. J. Psychiatr. : J. Ment. Sci.* 191 (5), 402–407. <https://doi.org/10.1192/bjp.bp.107.036129>.
- Zohoori, N., Savitz, D.A., 1997. Econometric approaches to epidemiologic data: relating endogeneity and unobserved heterogeneity to confounding. *Ann. Epidemiol.* 7 (4), 251–257. [https://doi.org/10.1016/S1047-2797\(97\)00023-9](https://doi.org/10.1016/S1047-2797(97)00023-9).