Published in final edited form as:
J Am Stat Assoc. 2023 ; 118(543): 1488–1499. doi:10.1080/01621459.2022.2157727.

Assessing the Most Vulnerable Subgroup to Type II Diabetes Associated with Statin Usage: Evidence from Electronic Health Record Data

Xinzhou Guo[*,a], **Waverly Wei**[*,b], **Molei Liu**[c], **Tianxi Cai**[d], **Chong Wu**[e], **Jingshen Wang**[b]

[a]Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong, Hong Kong

[b]Division of Biostatistics, UC Berkeley, Berkeley, CA

[c]Department of Biostatistics, Columbia Mailman School of Public Health, New York, NY

[d]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA

[e]Department of Biostatistics, MD Anderson Cancer Center, Houston, TX

Abstract

There have been increased concerns that the use of statins, one of the most commonly prescribed drugs for treating coronary artery disease, is potentially associated with the increased risk of new-onset Type II diabetes (T2D). Nevertheless, to date, there is no robust evidence supporting as to whether and what kind of populations are indeed vulnerable for developing T2D after taking statins. In this case study, leveraging the biobank and electronic health record data in the Partner Health System, we introduce a new data analysis pipeline and a novel statistical methodology that address existing limitations by (i) designing a rigorous causal framework that systematically examines the causal effects of statin usage on T2D risk in observational data, (ii) uncovering which patient subgroup is most vulnerable for developing T2D after taking statins, and (iii) assessing the replicability and statistical significance of the most vulnerable subgroup via a bootstrap calibration procedure. Our proposed approach delivers asymptotically sharp confidence intervals and debiased estimate for the treatment effect of the most vulnerable subgroup in the presence of high-dimensional covariates. With our proposed approach, we find that females with high T2D genetic risk are at the highest risk of developing T2D due to statin usage.

Keywords

Bootstrap; Causal inference; Debiased inference; Precision medicine

CONTACT Jingshen Wang jingshenwang@berkeley.edu Division of Biostatistics, UC Berkeley, Berkeley, CA.

[*]These authors contributed equally to this work and are alphabetically ordered.

Supplementary Materials
The online supplementary materials include implementation and theoretical details of the proposed method, additional simulation, case study, and sensitivity analysis results.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

1. Introduction

1.1. Motivation and Objectives

Coronary artery disease (CAD), a disease affecting the function of heart, is the leading cause of deaths worldwide (Skourtis et al. 2020). Over the past decades, efforts have been made in developing effective and safe drugs in preventing and treating CAD (Povsic et al. 2017). Among those novel agents, statins are perhaps the most commonly prescribed drugs due to their clear benefits in reducing the level of low–density lipoprotein (LDL) and subsequently lowering CAD risks through 3-hydroxy-3-methylglutaryl-coenzyme A reductase (HMGCR) inhibition (Nissen et al. 2005). Despite their clear benefits in reducing CAD risks, the use of statins is potentially associated with the increased risk of new-onset Type II diabetes (T2D) (Waters et al. 2013; Macedo et al. 2014; Mansi et al. 2015).

Although many studies have been conducted to investigate the potential side effects of statins in developing T2D, to date, there is still no robust evidence as to whether and on what kind of populations statin usage increases the risk of T2D. Take some frequently cited studies as examples. Rajpathak et al. (2009) find through meta-analysis that there is a small increase in T2D risk[1] associated with the use of statins, but this association is no longer significant after including the results from the WOSCOPS trial (Packard et al. 1998)—the first study investigating the association between T2D and statins. Recent studies also suggest that the effect of statins on T2D risk might be heterogeneous across different sub-populations and be more pronounced in certain subgroups defined by sex and baseline T2D genetic risk (Mora et al. 2010; Goodarzi et al. 2013). Nevertheless, existing studies may not lead to trustworthy findings in subgroups, because their statistical analyses either are conducted under randomized controlled trials (RCTs) with limited sample sizes whose results might not be generalizable beyond study population(Mora and Ridker 2006; Mora et al. 2010), or do not adjust for multiple comparisons issue when several candidate subgroups are under consideration (Waters et al. 2013).

In this case study, leveraging Partner Health System (PHS) biobank and electronic health record (EHR) data, we conduct subgroup analysis and assess the most vulnerable subgroup to T2D associated with statin usage from a novel biological perspective. We focus on the most vulnerable subgroup not only because pursuing the subgroup with the largest (adverse) treatment effect is a conventional practice in clinical studies (Naggara et al. 2011; Kubota et al. 2014), but also because a comprehensive understanding of the most vulnerable subgroup to T2D risk associated with statin usage could support precise clinical decisions and effective actions concerning the prescription of statins (Mora et al. 2010; Bornkamp et al. 2017; Guo and He 2020). Concretely, our study consists of three objectives: (i) designing a rigorous causal framework that systematically examines the causal effects of statin usage on T2D risk from observational data, (ii) uncovering which patient subgroup is most vulnerable for developing T2D after taking statins, and (iii) assessing the replicability and statistical significance of the most vulnerable subgroup via a bootstrap calibration procedure.

[1] Relative risk (RR): 1.13, 95% CI [1.03, 1.23]

1.2. Overview of Research Methods and Findings

To systematically examine the causal effect of statin usage on T2D risk from observational data, we propose a novel study design which not only circumvent common issues in RCTs but also alleviate the concern of unmeasured confounding bias.

On the one hand, while existing studies often investigate the adverse effects of stains on T2D risk in RCTs with limited sample sizes, our study design leverages the large PHS biobank with linked EHR data, providing robust evidence for assessing the adverse effect of statin usage. We extract and link genotype information together with diagnostics from consented subjects in the Partner Health System(PHS) biobank and EHR data, respectively. This leads to an EHR virtual cohort of 17,023 subjects, a much larger cohort than those from usual clinical trials, and 337 features; see Section 2.3 for detailed descriptions. Compared to study cohorts enrolled in RCTs, our study cohort can be a more representative sample of the general population (see Table 1 for the demographics of our study cohort).

On the other hand, while causal conclusions derived from observational studies can be susceptible to unmeasured confounding and reverse causation bias, our study design adopts a randomly inherited single nucleotide polymorphism (SNP), rs12916-T, as a *surrogate treatment variable* of statin usage to alleviate the concerns of those issues. rs12916-T is a reliable surrogate treatment variable because it resides in the *HMGCR* gene encoding the drug target of statins, and has been recently used as an unbiased, unconfounded proxy for pharmacological action on the target of statins (i.e., HMG-CoA reductase inhibition) (Swerdlow et al. 2015). Furthermore, adopting a randomly inherited genetic variant at conception as a surrogate for statin usage in EHRs allows us to establish a clear temporal precedence between the treatment and T2D onset (which is a prerequisite to concluding causality, see, e.g., Holland 1986). This avoids potential reverse causation issues. Lastly, because the surrogate treatment variable is naturally inherited, variables observed after birth are independent of rs12916-T and can at most be mediators that belong to a different causal pathway. We shall discuss the reasoning of using rs12916-T in more detail in Section 2.1.

Leveraging the above study design, we further conduct subgroup analysis and assess if subjects in different subgroups carrying the rs12916-T allele (i.e., taking statins) have heterogeneous risks at developing T2D, and to what extent the most vulnerable subgroup suffers from the side effect of statin usage. Inspired by study designs adopted in Mora et al. (2010) and Wang and Ware (2013), we divide our study cohort into six predefined candidate subgroups based on sex and baseline T2D genetic risk profiles (measured by the number of risk alleles of variants rs35011184-A and rs1800961-T each individual carries), and aim to uncover the most vulnerable patient subgroup to statin usage and assess the statistical significance of the most vulnerable subgroup. While numerous methods have been proposed in subgroup analysis for identifying subgroups (Lipkovich et al. 2011; Ma and Huang 2017) and testing subgroup homogeneity (Shen and He 2015; Fan, Song, and Lu 2017), in our case study, the primary objective is to make valid post-hoc inference on the most vulnerable subgroup.

The limitation of usual post-hoc inference on the most vulnerable subgroup is well recognized. Due to the winner's curse bias induced by multiple comparisons (Efron 2011), post-hoc inference often leads to false positive results (Thomas and Bornkamp 2017). Although several attempts have been made to address the winner's curse bias issue, existing procedures are either poorly grounded (Stallard, Todd, and Whitehead 2008; Rosenkranz 2016) or tend to be conservative (Hall and Miller 2010; Fuentes, Casella, and Wells 2018), as the latter is typically built on simultaneous inference aiming to control the family-wise error rate for all candidate subgroups. These conservative simultaneous inference procedures are usually undesirable in subgroup analysis, because they yield false negative discoveries and have inadequate power to confirm the most vulnerable subgroup (Magnusson and Turnbull 2013; Burke et al. 2015). In our context, because subgroup analysis needs to be conducted in observational studies (Lu et al. 2018; Yang et al. 2020), the problem becomes even more challenging as we need to take possibly high-dimensional confounders into account in assessing subgroup treatment effects. To address the above-mentioned issues, we provide new post-hoc inferential tools to help assess the efficacy of the most vulnerable subgroups from observational studies without having to resort to simultaneous inference methods, which are often too conservative to start with.

By applying the proposed method to our study cohort, we find that, although the overall adverse effect of statin usage on developing T2D is not significant, the female subgroup with high-genetic baseline T2D risk (more than two T2D risk alleles) is identified as the most vulnerable subgroup, meaning that with statin usage, this subgroup has the highest risk of developing T2D. The statistical significance of such a finding is also confirmed by the proposed bootstrap calibration method. In sum, our case study not only provides new evidence supporting the adverse effect of statin usage from a biological perspective, but also suggests that more caution should be taken when statins are prescribed, especially for females who are already at higher risk of developing T2D. The specific actions may include preventive treatments for diabetes and recommendations on lifestyle changes.

2. Data Description and Model Setup

2.1. Study Design

In our case study, since patients' statin use information is not available, we adopt the genetic variant (rs12916-T) as the surrogate treatment variable of statin usage. When the treatment indicator variable t = 1, this means that "the subject carries the variant rs12916-T." When t = 0, this means that "the subject does not carry the variant rs12916-T." We adopt this genetic variant as a surrogate treatment variable not only because statin usage information is not available in our EHR data, but also because carrying rs12916-T is a proxy for statin usage. The reason for adopting this proxy is due to the fact that rs12916-T, which resides in the *HMGCR* gene encoding the drug target of statins, has been recently adopted as an unbiased, unconfounded proxy for pharmacological action on the target of statins (Swerdlow et al. 2015). In other words, rs12916-T allele and statins are functionally equivalent in that they both lower LDL cholesterol level through HMG-CoA reductase inhibition. Concretely, Würtz et al. (2016) show that the metabolic changes (e.g., decreased LDL cholesterol level) associated with statin usage have been found to resemble the association between rs12916-T and the metabolic changes with $R^2 = 0.94$. Therefore, we believe that rs12916-T is a credible surrogate treatment variable of statin usage.

Furthermore, besides the absence of statin usage information in our EHR data, we shall argue that adopting rs12916-T as the surrogate treatment variable of statin usage invests our framework with two other key benefits.

First, because our study design ensures the temporal precedence between inheriting the variant rs12916-T (surrogate of statin use) and T2D onset, this lifts the concern of reverse causation in conducting causal inference from observational data. In particular, when studying the causal effect of statin use on T2D risk from observational data, one normally assumes that statin use causes the change in T2D risk (Pan et al. 2020). This suggests that only data collected from subjects who have recorded statin use status before T2D onsets can be used for credible casual analyses. Unfortunately, in our EHR data, it is impossible to establish the temporal precedence between statin use and T2D onset. Using a genetic variant (rs12916-T) as the surrogate treatment variable circumvents the above-mentioned issue. Because genetic variants are randomly inherited at conception, our study design guarantees that the cause (carrying rs12916-T [proxy for pharmacological action of statin use] or not) must occur before T2D onset. This clear temporal precedence makes the established causal relationships more plausible (see Figure 1).

Second, adopting rs12916-T as a surrogate treatment variable attenuates the concern of unmeasured confounding biases in observational data. Given rs12916-T is naturally inherited, variables observed after birth can at most be mediators that belong to a different causal pathway. Since we work with the causal pathway between the treatment and the outcome, that is, $t_i \rightarrow y_i$, conditional on such (unmeasured) mediators is not necessary. Thus, the unmeasured confounding bias issue is alleviated under our study design. To further robustify our causal conclusion and improve statistical estimation efficiency of the underlying causal effect, we include additional potential confounders that are obtained before birth, such as genetic variants associated with T2D and potentially associated with rs12916-T. We provide more discussions on causal pathways in Section H, supplementary materials.

Our outcome of interest is T2D status. We defer the technical details on defining T2D status from EHR data to Section 2.3. Given that the existing literature (Mora et al. 2010; Waters et al. 2013) has suggested that the treatment effect of statin usage on T2D risk could be heterogeneous, we conduct subgroup analysis to investigate the causal effect of T2D associated with statin usage in subgroups and the most vulnerable one in particular. Inspired by the study designs in Mora et al. (2010) and Waters et al. (2013) in which patient population is divided based on sex in the former and the number of T2D baseline risk factors in the latter, we divide our PHS study cohorts into six pre-defined candidate subgroups based on sex and baseline T2D genetic risk profiles. The baseline T2D genetic risk is measured by the number of copies of T2D risk allele of variants rs35011184-A and rs1800961-T each subject has. The more T2D risk alleles the subject has, the higher baseline genetic T2D risk the subject bears (Lango et al. 2008). We define “low-risk” as the total number of alleles = 0, “mid-risk” as the total number of alleles = 1, and “high-risk” as the total number of alleles ≥ 2. The six subgroups are thereby divided as (a) high-risk female; (b) mid-risk female; (c) low-risk female; (d) high-risk male; (e) mid-risk male, and (f) low-risk male.

Here, we consider pre-defined subgroups instead of post-hoc identified subgroups because pre-defined subgroups usually have clearer interpretability and could avoid the bias induced by data-adaptive subgroup identification procedure, while post-hoc identified subgroups are often adopted when there is no prior information on the segregation of study population (Lipkovich et al. 2011; Ma and Huang 2017). In our setting, because previous studies (Mora et al. 2010; Waters et al. 2013) suggest that T2D risk might have differential effects across sex and baseline T2D genetic profiles, pre-defined subgroups are more suitable for the present case study. In Section G, supplementary materials, we compare the pre-defined subgroups and post-hoc identified subgroups based on our EHR data, and we find that the post-hoc identified subgroups resemble the pre-defined subgroups adopted in our case study.

2.2. Model Setup

We work with the following sparse logistic regression model:

$$\operatorname{logit}\left\{\mathbb{P}(\mathrm{y} = 1 \mid \mathrm{z}, \mathrm{x})\right\} = \mathrm{z}^{\top}\beta + \mathrm{x}^{\top}\gamma, \quad \|\gamma\|_0 \ll p. \tag{1}$$

Here, y is the observed binary outcome representing the T2D status. $\mathrm{z} \in \mathbb{R}^{p_1 \times n}$ includes variables representing interactions between the treatment variable and all the six subgroup indicator variables. $\mathrm{x} \in \mathbb{R}^{p_2 \times n}$ contains 336 covariates and an intercept (hence, $p_2 = 337$). The 336 covariates contain five subgroup indicator variables (the sixth subgroup, low-risk male, indicator variable is dropped to avoid collinearity) and 331 potential confounders (including race and age as baseline characteristics, and 329 SNPs associated with T2D related factors accounting for potential confounding issues). Note that we do not include the treatment variable as a covariate because including it causes collinearity issues. All observed covariates are obtained from Partner Health System biobank.

Following the above setup, $\beta \in \mathbb{R}^{p_1}$ represents subgroup causal effects on the scale of log odds ratio (OR) (hence, $p_1 = 6$). More concretely, under Model (1), $\beta = (\log\alpha_1, \ldots, \log\alpha_6)$ with $\log\alpha_1$ representing the log odds ratio of subgroup j, for $j = 1, \ldots, 6$. Following the Neyman-Rubin causal model and our current study design, we provide rigorous causal identification results to justify why the model parameterization in (1) enables us to estimate the heterogeneous causal effects in the pre-defined subgroups. This theoretical justification is provided in Section E, supplementary materials. We further assume that $\gamma \in \mathbb{R}^{p_2}$ is a sparse vector with the support set M_0. The sparsity assumption not only provides a parsimonious explanation of the data but also carries our prior belief that not every genetic variant is predictive of the outcome as demonstrated in Section 2.3.

2.3. Data Description and Exploration

Following our study design described in the previous section, we extract and link genotype information and diagnostics from consented subjects in the PHS biobank and EHR data respectively. Our data involve a much larger cohort than those from usual clinical trials, $n = 17,023$ subjects each with $p = 337$ features; see Table 1 for data summary.

Recall that the covariates contain age, race, subgroup indicators, and genetic variants associated with T2D related factors (including LDL, high density lipoprotein and obesity). As for the definition of the outcome, since the diagnostic billing code for T2D has limited specificity in classifying the true T2D status, we define the T2D status based on a previously validated multimodal automated phenotyping (MAP) algorithm (Liao et al. 2019). The area under the ROC curve (AUC) of MAP's risk prediction score for classifying true T2D status is 0.99, and the specificity and sensitivity of its classifier are 0.97 and 0.92, respectively. These suggest that the MAP classifier of T2D can be reliably used to define the T2D outcome. Among our study cohort, MAP classifies 2565 subjects having T2D. There is no missing data issue in our case study for two reasons. First, we leverage the large PHS biobank with linked EHR data, thus, the genetic profiles and baseline covariates are non-missing. Second, we adopt surrogate outcomes, thus, we do not encounter any missing outcomes.

To explore the association between statin usage and T2D risk, we report preliminary data exploration results from a "full" logistic regression model for y against the treatment t and the covariates x in Table 2; that is, $y \sim \mathrm{t} + \mathrm{x}$. There, although a modest association is found between carrying the rs12916-T variant and T2D status in the overall study cohort, unlike the results in Swerdlow et al. (2015), this association is not statistically significant. Moreover, our analysis reports only 16 regression coefficients having p-values < 0.05, suggesting that the logistic regression coefficient vector for the covariates is likely to be sparse. Because the overall treatment effect is marginal (estimated marginal treatment effect equals 0.04 with p-value 0.35), this motivates us to conduct subgroup analysis to further investigate the subgroup causal effect of statin usage on T2D risk.

2.4. Challenges in Statistical Inference

Because our goal is to assess the patient subgroup most vulnerable for developing T2D, our methodological development, hence, centers around delivering valid inference (accurate point estimate and valid confidence interval) on the effect size of maximal regression coefficient $\beta_{\max} = \max_{j \in \{1, \ldots, p_1\}} \beta_j$ in Model (1). We focus on $\beta_{\max}$ instead of $|\beta|_{\max}$ for the following reason. The logistic regression coefficient β represents the log odds ratio, thus, each regression coefficient is a number ranging from $-\infty$ to ∞. A larger β indicates a higher T2D risk associated with statin usage. If we use $|\beta|$, a larger $|\beta|$ might no longer measure the adverse effect of statin usage on T2D risk, implying that $|\beta|_{\max}$ represents the adverse effect of either the most vulnerable subgroup or the least vulnerable subgroup to statin usage. Because $\beta_{\max}$ has clearer interpretation than $|\beta|_{\max}$, we focus on $\beta_{\max}$ instead of $|\beta|_{\max}$ in this article.

In the presence of high-dimensional covariates as described in Section 2.3, finding an accurate point estimate and conducting inference on $\beta_{\max}$ can be a challenging task, due to the presence of regularization and winner's curse biases. The regularization bias occurs whenever penalization approaches are adopted to select a smaller working model to enhance the estimation efficiency of β in the presence of sparsity (Hong, Kuffner, and Martin 2018; Wang, He, and Xu 2019). The winner's curse bias occurs whenever we use a simple sample-analogue $\widehat{\beta}_{\max} = \max_{j \in \{1, \ldots, p_1\}} \widehat{\beta}_j$ to estimate the true maximum effect $\beta_{\max}$. A sample average estimate for $\beta_{\max}$ overestimates the parameter because, even if $\widehat{\beta}$. follows normal distribution centering at β, $\widehat{\beta}_{\max}$ will follow a skewed-normal distribution and will not center at $\beta_{\max}$ (Nadarajah and Kotz 2008; Guo and He 2020). Such an overestimation phenomenon is well-recognized in post-hoc subgroup analysis (see, e.g., Zöllner and Pritchard 2007; Cook et al. 2014). While several approaches have been proposed to address the regularization bias issue (Zhang and Zhang 2014; Li 2020), and provide valid inference on a single regression coefficient, these methods cannot account for the winner's curse bias. As for the winner's curse bias, existing methods are mostly made for low-dimensional data and are not directly applicable to observational data with high dimensional covariates in this case study (Bornkamp et al. 2017; Guo and He 2020). While Guo et al. (2021) proposes bootstrap-based approaches to simultaneously address the regularization and winner's curse bias issues in high dimensional linear models, we broaden its validity by providing a bootstrap procedure that is asymptotically valid for high dimensional logistic regression estimators with rigorous statistical guarantees. To our knowledge, debiasing procedures that simultaneously remove the regularization bias and winner's curse bias in high dimensional nonlinear models have been lacking. Technical discussions on these bias issues are deferred to the Section A, supplementary materials, and we demonstrate the winner's curse bias and regularization bias in estimating $\beta_{\max}$ within a simple simulated example.

Example 1 (Winner's curse bias and regularization bias in estimating $\beta_{\max}$).—

We use two widely adopted procedures to estimate β: (a) Lasso for generalized linear models (GLM) (Park and Hastie 2007), which estimates β with $\left(\widehat{\beta}_{\text{GLasso}}^{\top}, \widehat{\gamma}_{\text{GLasso}}^{\top}\right)^{\top}$ obtained from the ℓ_1-penalized logistic regression program without any adjustments, and (b) Refitted GLM Lasso, which estimates β by refitting the logistic regression model based on the covariates in the support set of $\left(\widehat{\beta}_{\text{GLasso}}^{\top}, \widehat{\gamma}_{\text{GLasso}}^{\top}\right)^{\top}$. As a benchmark, we also report the performance of the oracle estimator $\left(\widehat{\beta}_{\text{Oracle}}^{\top}, \widehat{\gamma}_{\text{Oracle}}^{\top}\right)^{\top}$ which pretends the true support set of γ is known and is estimated by refitting the logistic regression model with the true support set. $\beta_{\max}$ is then estimated in a two-step procedure: One first obtains an estimate $\widehat{\beta}$ and then estimates $\beta_{\max}$ by taking the maximum, that is $\max\left\{\widehat{\beta}_1, \ldots, \widehat{\beta}_{p_1}\right\}$. To mimic the causal relationship in this case study, we generate Monte Carlo samples with $\mathrm{t}_i \sim \text{Bernoulli}(0.5)$ independent of the covariate $\mathrm{w}_i \sim N(0, \Sigma)$, where $\Sigma = (\Sigma_{jk})_{j,k=1}^{p-6}$ and $\Sigma_{jk} = 0.5^{|j-k|}$ for $i = 1, \ldots, n$. We then generate $\mathrm{x}_{ij} = \mathbb{1}(\mathrm{w}_{ij} > 0)$ for 1 $1 \leq j \leq p - 6$, and $z_{il} = t_i x_{il}$, $l = 1, \ldots, 6$. y_i is generated following Model (1). We set the sample size $n = 1000$ and the dimension $p = 200$, and set $\gamma = (1, 1, 0, \ldots) \in \mathbb{R}^{p-6}$. For the first simulation, we set the coefficients $\beta = (0.5, 0.5, 0, 0, 0, 0)^{\top}$ and vary the value of tuning parameter λ to illustrate how the winner's curse bias could invalidate the inference as the winner's curse bias is the most severe when the two largest β's are equal (Nadarajah and Kotz 2008; Guo and He 2020). The results are shown in Figure 2(A). For the second simulation, to demonstrate how the winner's curse bias changes with respect to the distance between the largest and the second largest components in β, that is, $\beta_{(1)} - \beta_{(2)}$, we fix $\log\lambda = -2.5$ for illustration, set the coefficients $\beta = (\beta_{\max}, 0.5, 0, 0, 0, 0)^{\top}$, where $\beta_{\max} \in \{0.5, 0.61, 0.72, \ldots, 1.5\}$, and plot the $\sqrt{n}$-scaled bias with respect to various $\beta_{(1)} - \beta_{(2)}$ values, where $\beta_{(1)}$ is equivalent to $\beta_{\max}$. The results are presented in Figure 2(B). In Figure 2, we report the root-n scaled bias based on 500 Monte Carlo samples under the two settings respectively.

From the results in Figure 2, we observe that all three estimators are biased. Although $\hat{\beta}_{\text{Oracle}}$ is a consistent estimator of β, its maximum is not centered around $\beta_{\max}$. Following some explicit evidence given in Nadarajah and Kotz (2008), $\hat{\beta}_{\max}$ is usually biased upward for estimating $\beta_{\max}$, we thus conjecture that the residual bias in the maximal of the oracle estimator $\hat{\beta}_{\text{Oracle,max}}$ is caused by the winner's curse bias issue. We further observe that the magnitude of the winner's curse bias decreases as the distance between $\beta_{(1)}$ and $\beta_{(2)}$ increases (as seen in Figure 2(B)), suggesting that the winner's curse bias might not be a severe concern if $\beta_{(1)}$ and $\beta_{(2)}$ are far apart. As β is unknown a priori, inference procedure without adjusting for the winner's curse bias may not be valid in practice. On the top of the winner's curse bias issues, the GLM Lasso and the refitted estimators suffer from the regularization bias and hence are also not correctly centered around $\beta_{\max}$, unless in some special cases where the regularization bias and the winner's curse bias cancel out.

To simultaneously adjust for the winner's curse bias and the regularization bias without knowing the underlying true parameters, in what follows, we propose an inferential framework that produces a bias-reduced estimate as well as a valid confidence interval of $\beta_{\max}$.

3. Methodology

We start with describing an estimation strategy of β that resolves the regularization bias induced by model selection and helps addressing our research objectives (ii) and (iii) discussed in Section 1.1. Regularization bias arises when the selected model is either over-fitted or under-fitted; see detailed discussion provided in Section A, supplementary materials. While the risk of under-fitting can be mitigated by aiming for a larger model for parameter estimation, we resolve the issue of over-fitting by sample splitting. Sample splitting divides a sample into two parts: The first part of the sample is used for model selection and the remaining part is used for estimation based on the selected model. When γ is sparse and a larger model is selected based on the first half of the sample, we expect refitted GLM estimator on the second part of the sample to be free of significant bias. Nevertheless, sample splitting provides debiased estimator of β at a cost of increased variability, because only a part of the sample is used for estimation. To minimize this efficiency loss due to sample splitting, we consider the method of repeated sample splitting (R-Split) that averages different estimates of β across different splits. Our strategy, in a spirit similar to bagging and ensemble algorithms in machine learning, helps to stabilize and improve the accuracy of the estimated β in a subsample.

Step 1 (Repeated sample splitting that accounts for the regularization bias) For $b \leftarrow 1$ to B_1: (1) Randomly split the sample $\{(y_i, x_i, z_i)\}_{i=1}^{n}$ into two subsamples: a subsample T_1 of size n_1

and a subsample T_2 of size $n_2 = n - n_1$; (2) select a model $\widehat{\mathrm{M}}_b$ to predict y based on T_1; (3) refit the selected model with the data in T_2 to estimate β_b and γ_b via logistic regression:

$$\left(\widehat{\beta}_b^{\top}, \widehat{\gamma}_b^{\top}\right)^{\top} = \operatorname{argmin}\left\{\sum_{l \in T_2} \left(\mathrm{y}_l \cdot \left(\mathrm{z}_l^{\top}\beta + \mathrm{x}_{l,\widehat{\mathrm{M}}_b}^{\top}\gamma\right)\right.\right.$$
$$\left.\left. - \log\left(1 + \exp\left(\mathrm{z}_l^{\top}\beta + \mathrm{x}_{l,\widehat{\mathbb{M}}_b}^{\top}\gamma\right)\right)\right)\right\}.$$

(4) obtain the R-Split estimate: $\tilde{\beta} = \frac{1}{B_1}\sum_{b=1}^{B_1}\widehat{\beta}_b$.

In this step, any reasonable model selection procedures may be used and the choice of model size is subjective, but the selected model needs to be large enough for the under-fitting bias to be negligible. In our simulation and case study, we use GLM Lasso for model selection (Friedman et al. 2017) and choose the model size from cross-validation (see Section B, supplementary materials for detailed description). The choice of splits B_1 needs to be sufficiently large so that the R-Split estimator $\tilde{\beta}$ has a tractable asymptotic distribution. Under appropriate regularity conditions, we show that $\tilde{\beta}$ converges to a normal distribution centered around β at a root-n rate (statistical justification is provided in the Section C.2, supplementary materials).

As $\tilde{\beta}$ provides an accurate estimate of β we use $\tilde{\beta}$ to address our research objectives (ii) and (iii). In particular, the subgroup with the largest coefficient, $\operatorname{argmax}_{j \in [p_1]}\tilde{\beta}_j$, is most vulnerable for developing T2D after taking statins. However, due to the winner's curse bias, simply relying on $\tilde{\beta}$ will not lead to valid inference on $\beta_{\max}$, and we need a second step to address objective (iii). Built upon an accurate estimate of β, we store an inverse Hessian matrix for the later bootstrap calibration to adjust for the winner's curse bias:

$$\widetilde{\Gamma}_n = \frac{1}{B_1}\sum_{b=1}^{B_1} \mathrm{I}_z\left(\frac{1}{n_1}\sum_{i \in T_{2,b}} f_{i,b}\begin{pmatrix}\mathrm{z}_i \\ \mathrm{x}_{i,\widehat{\mathbb{M}}_b}\end{pmatrix}\left(\mathrm{z}_i^{\top}, \mathrm{x}_{i,\widehat{\mathbb{M}}_b}^{\top}\right)^{\top}\right)^{-1}\mathrm{I}_{\widehat{\mathbb{M}}_b},$$

where $f_{i,b} = \operatorname{expit}'\left(\mathrm{z}_i^{\top}\widehat{\beta}_b + \mathrm{x}_{i,\widehat{\mathbb{M}}_b}^{\top}\widehat{\gamma}_b\right)$. Its benefits will be apparent in the following step:

Step 2 (Calibrated bootstrap that accounts for the winner's curse bias) For $b \leftarrow 1$ to B_2: generate bootstrap replicate $\tilde{\beta}^*$ from:

$$\tilde{\beta}^* = \tilde{\beta} + \widetilde{\Gamma}_n \cdot \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}\mathrm{z}_i \\ \mathrm{x}_i\end{pmatrix}\upsilon_i^*, \tag{2}$$

where $\upsilon_i^* = u_i\widehat{\upsilon}_i$ is the permuted GLasso residual, $\widehat{\upsilon}_i = \mathrm{y}_i - \operatorname{expit}\left(\mathrm{z}_i^{\top}\widehat{\beta}_{\mathrm{GLasso}} + \mathrm{x}_i^{\top}\widehat{\gamma}_{\mathrm{GLasso}}\right)$ Then recalibrate bootstrap statistics via

$$T_b^* = \max_{j \in [p1]}\left(\tilde{\beta}_j^* + \tilde{c}_j(r)\right) - \tilde{\beta}_{\max},$$
$$\tilde{c}_j(r) = \left(1 - n^{r - 0.5}\right)\left(\tilde{\beta}_{\max} - \tilde{\beta}_j\right), \text{ where} r \in (0, 0.5).$$

In this step, rather than adopting the simple bootstrap statistics $\max_{j \in [p1]}\tilde{\beta}_j^* - \tilde{\beta}_{\max}$ to make inference on $\beta_{\max}$, we make an adjustment to each coordinate of $\tilde{\beta}^*$ by the amount $\tilde{c}_j(r)$. This is because just as $\tilde{\beta}_{\max}$ is a biased estimator of $\beta_{\max}$, the simple bootstrap statistics $\max_{j \in [p1]}\tilde{\beta}_j^* - \tilde{\beta}_{\max}$ is also not centered at $\tilde{\beta}_{\max}$. The amount of adjustment $\tilde{c}_j(r)$ is large when $\tilde{\beta}_j$ is small, and is small when $\tilde{\beta}_j$ is large. By adding the correction term $\tilde{c}_j(r)$, under certain regularity conditions, the distributions of $\sqrt{n}\left(\tilde{\beta}^*_{\text{modified;max}} - \tilde{\beta}_{\max}\right)$ and $\sqrt{n}\left(\tilde{\beta}_{\max} - \beta_{\max}\right)$ are asymptotically equivalent, implying that our proposed method adjusts for the winner's curse bias and the regularization bias simultaneously, where $\tilde{\beta}^*_{\text{modified;max}} = \max_{j \in [p1]}\left(\tilde{\beta}_j^* + \tilde{c}_j(r)\right)$. We relegate the theoretical details of this bootstrap calibration procedure in Section C, supplementary materials. Note that $r \in (0, 0.5)$ is a positive tuning parameter (see Section B, supplementary materials for its data adaptive choice).

At this point, we note that our procedure adopts wild bootstrap to construct bootstrapped statistics of the R-Split estimate $\tilde{\beta}$. The wild bootstrap procedure adopted here is not only computationally efficient in high dimensions, as the Hessian matrix remains unchanged across different bootstrap samples, but also provably consistent in our problem setup. Furthermore, Dezeure, Bühlmann, and Zhang (2017) shows that wild bootstrap can be more versatile than other residual bootstrap methods because it correctly captures the asymptotic variance for various settings. With the help of a valid bootstrap calibration procedure in replicating $\tilde{\beta}_{\max}$, we are now ready to propose our final step that constructs confidence intervals and debiased estimate for $\beta_{\max}$:

Step 3 (Bias-reduced $\tilde{\beta}_{\max}$ and sharp confidence interval) The level-α two-sided confidence interval for $\left[\tilde{\beta}_{\max} - Q_{T_b^*}(\alpha/2), \tilde{\beta}_{\max} + Q_{T_b^*}(\alpha/2)\right)$, and a bias-reduced estimate for $\dot{\beta}_{\max}$ is $\beta_{\max} - \frac{1}{B_2}\sum_{b=1}^{B_2} T_b^*$.

4. Theoretical and Empirical Justification

In this section, we provide theoretical justifications of the proposed bootstrap-assisted R-Split estimator along with a simple power analysis, where we demonstrate that our approach not only has rigorous theoretical guarantee but also shows high statistical detection power. We then examine the performance of the proposed method through simulation studies.

4.1. Theoretical Investigation and a Power Analysis

The following theorem confirms that the asymptotic distribution of $\sqrt{n}\left(\tilde{\beta}_{\text{modified;max}} - \tilde{\beta}_{\max}\right)$ converges to $\sqrt{n}\left(\tilde{\beta}_{\max} - \beta_{\max}\right)$. This suggests that the proposed confidence interval constructed in Step 3 of Section 3 is "asymptotically sharp," meaning that it achieves the exact nominal level as the sample size goes to infinity. This distinguishes the proposed procedure from other conservative methods made for subgroup analysis (e.g., Hall and Miller 2010;

Fuentes, Casella, and Wells 2018). The proof of Theorem 1 is provided in the Section C.3, supplementary materials. To simplify presentation, we relegate regularity assumptions to the Section C.1, supplementary materials.

Theorem 1.—Under Assumptions 1–9 given in Section C.1, supplementary materials, when p_1 is a fixed number, the modified bootstrap maximum treatment effect estimator, $\tilde{\beta}^{*}_{\text{modified;max}} = \max_{j \in [p_1]}\left(\tilde{\beta}^{*}_{j} + \tilde{c}_j(r)\right)$, satisfies:

$$\sup_{c \in \mathbb{R}} \mid \mathbb{P}\left(\sqrt{n}\left(\tilde{\beta}_{\max} - \beta_{\max}\right) \le c\right) \\ - \mathbb{P}^{*}\left(\sqrt{n}\left(\tilde{\beta}^{*}_{\text{modified;max}} - \tilde{\beta}_{\max}\right) \le c\right) \mid \ = o_p(1) .$$

The above theoretical result has two direct implications. On the one hand, as the proposed bootstrap calibration strategy successfully replicates the distribution of $\sqrt{n}\left(\tilde{\beta}_{\max} - \beta_{\max}\right)$, our bias-reduced estimator discussed in the Step 3 of Section 3 simultaneously removes the regularization bias and the winner's curse bias in $\tilde{\beta}_{\max}$. On the other hand, although simultaneous inference also delivers valid inference on $\beta_{\max}$ with strict Type-I error rate control, our proposal delivers valid inference on $\beta_{\max}$ without sacrificing the statistical power. This property is more desirable in our problem setup as we aim to look for the subgroup with the most severe side effect of statin usage while simultaneous methods often lead to overly conservative conclusions for this purpose.

To further demonstrate the merit of constructing an asymptotically sharp confidence interval for $\beta_{\max}$ and the benefit of conducting variable selection in finite samples, we compare statistical power for testing the null hypothesis H_0: $\beta_{\max} = 0$ for four procedures: (a) the proposed bootstrap-assisted R-Split, (b) R-Split with simultaneous confidence intervals, (c) the proposed bootstrap-assisted logistic regression, and (d) the desparsified Lasso estimator discussed in Zhang and Zhang (2014) with simultaneous confidence interval (Dezeure, Bühlmann, and Zhang 2017; Fuentes, Casella, and Wells 2018). We follow the same simulation setup as the first simulation setup in Example 1. The tuning parameter is fixed at $r = 0.15$ for simplicity. For R-Split, we choose the model size via cross-validation (see, Section B, supplementary materials) with a minimal model size equals 3 and a maximal model size equals 10.

From Figure 3, we observe that all considered approaches control the Type-I error rate at the nominal level when $\beta_{\max} = 0$. The bootstrap-assisted R-Split has the highest detection power over a range of $\beta_{\max}$ among all considered procedures. The bootstrap-assisted logistic regression has the lowest detection power, which demonstrates the necessity of conducting variable selection to screen out irrelevant predictors. As we have expected, both the R-Split method with simultaneous confidence interval and the desparsified Lasso with simultaneous confidence interval do not retain sufficient statistical power to detect subgroup treatment effect heterogeneity.

4.2. Simulation Studies

In this section, we consider various simulation designs to demonstrate the merit of our proposal. There are three main takeaways from this simulation. First, our proposed bootstrap calibration procedure provides confidence intervals with nominal coverage probabilities of $\beta_{\max}$ in finite samples. Second, R-Split based methods provide more accurate point estimates and shorter confidence intervals than the logistic regression based approaches without variable selection. Third, the bootstrap-assisted methods have higher statistical efficiency (shorter confidence intervals) compared to the simultaneous methods.

We generate Monte Carlo samples from the following model:

$$\operatorname{logit}\{\mathbb{P}(\mathrm{y}_i = 1 \mid \mathrm{z}_i, \mathrm{x}_i)\} = \mathrm{z}_i^\top \beta + \mathrm{x}_i^\top \gamma, \quad i = 1, \ldots, n,$$

with $n = 2000$. We consider two cases for β: (a) heterogeneous case with $\beta = (0, \ldots, 0, 1)^\top \in \mathbb{R}^{p_1}$, meaning that there exists subgroup treatment effect heterogeneity and only one subgroup singles out; and (b) spurious heterogeneous case with $\beta = (0, \ldots, 0, 0) \in \mathbb{R}^{p_1}$, meaning that there is no subgroup with significant treatment effect in the population. We set $\gamma = (1, 1, 1, 1, 0, \ldots, 0) \in \mathbb{R}^{p_2}$. In all considered simulation designs, we set $p_1 \in \{4, 10\}$. We consider the case with $(n, p_2) = (2{,}000{,}150)$ for logistic regression, R-Split, and the desparsified Lasso (Zhang and Zhang 2014), and consider the case with $(n, p_2) = (2000, 500)$ for R-Split and the desparsified Lasso, since logistic regression tends to provide inconsistent estimates in moderately high dimensions (Sur and Candès 2019). In each simulation design, we first take the maximum of estimated subgroup treatment effects, that is, $\widehat{\beta}_{\max} = \max_{j = 1, \ldots, p1} \widehat{\beta}_j$, in each Monte Carlo sample to mimic the subgroup selection procedure adopted in practice, and then we take the average across different Monte Carlo samples to calculate the winner's curse bias.

As for the covariates design, we generate z_i and x_i from

$$\mathrm{z}_{ij} \sim \operatorname{Bernoulli}\left(\frac{\exp(\mathrm{x}_{i, 2j-1} + \mathrm{x}_{i, 2j})}{1 + \exp(\mathrm{x}_{i, 2j-1} + \mathrm{x}_{i, 2j})}\right), \quad j = 1, \ldots, p_1,$$

where $\mathrm{x}_i \sim N(0, \Sigma)$ with $\Sigma_{ij} = 0.5^{|i-j|}$. We compare the finite sample performance of the proposed bootstrap-assisted R-Split and the bootstrap-assisted logistic regression with two benchmark methods: (a) a naive method with no bootstrap calibration, which directly uses the estimated maximum coefficient to estimate $\beta_{\max}$ and (b) the simultaneous method as discussed in Dezeure, Bühlmann, and Zhang (2017) and Fuentes, Casella, and Wells (2018). For the desparsified Lasso (Zhang and Zhang 2014), we only consider the above-mentioned two benchmark methods: the naïve method and the simultaneous method (without bootstrap calibration). For the R-Split method, we choose the model size via cross-validated GLM Lasso with a minimal model size equals 3. We report the coverage probability, the $\sqrt{n}$ scaled confidence interval length and the $\sqrt{n}$ scaled Monte Carlo bias along with their standard errors based on 1000 Monte Carlo samples in Table 3.

Comparing the bootstrap-assisted methods with the naive methods, we observe that the bootstrap-assisted methods have nominal-level coverage, while the naive methods are biased and under-covered. This comparison verifies the theoretical results in Section 4.1 that the proposed bootstrap calibration successfully reduces the winner's curse bias.

Comparing the bootstrap-assisted methods with the simultaneous methods, we find that although simultaneous methods have higher coverage probabilities, the confidence intervals are rather long, implying that simultaneous methods are overly conservative. While our proposed inferential framework reaches the nominal-level coverage probabilities and has shorter confidence intervals leading to asymptotically sharp inference.

The comparison between the bootstrap-assisted R-Split with the bootstrap-assisted logistic regression shows that the latter has larger biases and lower coverage probabilities. The bootstrap-assisted logistic regression has undesirable performance because logistic regression yields biased estimates in moderately high dimensions (Sur and Candès 2019). This comparison reveals the benefit of conducting variable selection when γ is sparse and the dimension of covariates is large, and it confirms that R-Split alleviates the regularization bias issue. Comparing R-Split with the desparsified Lasso, in line with our earlier conjecture in Section 4.1, we observe that the desparsified Lasso approach has wider confidence intervals than those obtained by R-Split and tends to provide conservative inference.

This simulation study verifies that our proposed inferential framework not only achieves nominal coverage probabilities, but also mitigates the regularization and winner's curse biases. Thus, the proposed inferential framework is sensible to consider for our case study.

5. Case Study

5.1. Case Study Results

In this section, we investigate the adverse effect of statin usage in our pre-specified six subgroups divided by sex and T2D genetic risk using the data introduced in Section 2.3. We compare the results from three methods: (a) repeated sample splitting (R-Split) without bootstrap calibration, (b) R-Split based on the simultaneous method discussed in Dezeure, Bühlmann, and Zhang (2017), and (c) the proposed bootstrap-assisted R-Split. We summarize our real data analyses results in Table 4, in which we have reported the estimated subgroup treatment effects from R-Split along with their p-values and two-sided confidence intervals, adjusted p-values to account for the multiple comparisons issue with simultaneous method and Bonferroni correction, and bootstrap calibrated p-values for the subgroup with the largest treatment effect. The results with one-sided confidence lower bounds are summarized in Section F, supplementary materials.

From Table 4, the results of the R-Split estimator without bootstrap calibration not only indicate that the treatment effect of statins tends to vary across different subgroups, but also suggest that the high-genetic-risk female subgroup is the most vulnerable group for developing T2D with estimated log-odds ratio 0.41, 95% two-sided confidence interval 0.04–0.78 ($\mathrm{OR} = 1.04 - 2.18$) with p-value 0.030. For males with various genetic risk levels and females with lower T2D genetic risk, the adverse effects of statin usage are not significant based on R-Split without bootstrap calibration. The treatment effect in the overall study cohort is slightly positive but is not significant, which is in-line with our expectation from the preliminary analysis in Section 2.3.

Although the estimates and confidence intervals from the R-Split without bootstrap calibration suggest that taking statins causes the increased risk of developing T2D for the most vulnerable subgroup, the statistical significance of this finding is unclear since R-Split is implemented without bootstrap calibration and can not address the multiple comparisons issue as illustrated in Section 4.2. After accounting for the multiple comparisons issue through conservative procedures including the simultaneous method or Bonferroni correction, the p-values for the female high-risk group are no longer significant, seemingly suggesting that our data do not provide enough evidence to claim the existence of the adverse effect of statin usage in the female high-risk subgroup. This might be due to the fact that both the simultaneous method and Bonferroni correction are rather conservative and tend to provide false negative discoveries. Fortunately, our proposed bootstrap assisted R-Split procedure directly conducts inference on the most vulnerable group, and our results suggest that among high-genetic-risk female patients, the odds of developing T2D after taking statins are 1.42 times the odds of developing T2D for the patients without taking statins (p-value 0.037 for two-sided test).

Our findings are in-line with reported results in existing clinical studies. For example, Mora et al. (2010) suggest that statin usage incurs a larger T2D risk increment on females than on males, and Waters et al. (2013) suggest that statins only significantly increase the risk of T2D on those with at least three out of four common T2D risk factors at baseline.[2] Compared with the existing studies, our findings provide more robust evidence with the new data analysis pipeline built under the causal inference framework. Our data analysis pipeline addresses several limitations of existing studies; in particular, limited sample size and multiple comparisons issue. Moreover, compared to existing studies, our findings provide a more biologically driven depiction of statins' heterogeneous adverse effect, which can further support effective and precise clinical decisions and actions concerning the prescription of statins. Our study further demonstrates that in practice, the genetic profiles could assist T2D prevention of statin receivers to improve the quality of clinical practices.

5.2. Sensitivity Analysis

A major concern in observational studies is the bias induced by unmeasured confounding, meaning that some unmeasured factors that are associated with both the treatment and the outcome may explain away the estimated causal effects (Robins, Rotnitzky, and Scharfstein 2000). To evaluate the validity of causal conclusions derived from our real data analyses, we conduct sensitivity analyses with the E-value method. The E-value method computes the minimal strength of an unmeasured confounder needed to explain away the estimated causal effect (VanderWeele and Ding 2017). Practitioners could then evaluate if there exists such an unmeasured confounder with the strength quantified by the E-value. A larger E-value implies that the unmeasured confounder needs to have a stronger association with the

[2] The risk factors used by Waters et al. (2013) include high fasting blood glucose, history of hypertension, high body mass index, and high fasting triglycerides.

outcome and the treatment in order to explain away the causal evidence. The E-values for our estimated subgroup causal effects are summarized in Table 5. Table 5 shows that the E-value in the high-risk female group is 2.38, which implies that only when an unmeasured confounder is associated with both the treatment and the outcome 2.38 times stronger than the measured confounders could the estimated causal effect be explained away. According to a meta-study on E-value applications, most computed E-values from existing literature are below 2.0 (Ioannidis et al. 2019). In sum, the results from Table 5 imply that the causal evidence collected from our data is reasonably robust against the unmeasured confounding issues.

Given that our outcome is an error-prone surrogate of the true disease status, we also conduct a sensitivity analysis regarding the potential misspecification of the logistic regression model for the true EHR disease status against the covariates. Due to page limit, the design and results of this sensitivity analysis are deferred to Section I, supplementary materials.

6. Discussion

In this case study, we investigate the T2D risk associated with statin usage in the most vulnerable subgroup. To overcome the limitations of existing studies and to generate trustworthy evidence, we introduce a rigorous study design under the causal inference framework and based on the EHR and biobank data from the Partner Health System. Built on this study design, we find that although the adverse effect of statin usage for developing T2D is marginal for the overall study cohort, taking statins significantly increases the risk of developing T2D for female patients with high genetic predisposition to T2D. We also recognize that our study design has two limitations. First, as the treatment variable is defined as if the subject carries the rs12916-T allele or not, we can only investigate the causal effect of taking statins on T2D risk but not the *dosage effect* of statins. Second, the definition of T2D status is based on a previously validated Multimodal Automated Phenotyping (MAP) algorithm (Liao et al. 2019). Although the MAP classifier of T2D can be reliably used to define the T2D outcome, generalizing the current study findings still warrants further confirmation from clinical trials.

While the objective of this case study is to make inference on the most vulnerable subgroup, a natural question to ask is whether statin usage will significantly increase the T2D risk for other vulnerable subgroups. To answer this question, we need to develop appropriate statistical tools to mitigate the regularization bias and winner's curse bias for other most vulnerable subgroups as well. Take the subgroup with the second largest treatment effect as an example, our proposed method might be extended to address the bias issues by appropriately modifying the correction term $\tilde{c}_j(r)$ to capture the distance between the second largest coefficient and the jth largest coefficient. We shall leave the rigorous methodology development for making valid inference on the other subgroups to future research.

This case study considers pre-defined candidate subgroups. While predefined subgroups are suitable in our case study (as discussed in Section 2.1), extending the proposed methodology to data-adaptively identified subgroups warrants future research. Data-adaptive subgroup identification approaches include, for example, varying coefficient model based (Chen and He 2018), regression tree based (Lipkovich et al. 2011), and fused Lasso based (Ma and Huang 2017) methods. When working with data-adaptively identified subgroups, one needs to not only adjust for the regularization and winner's curse bias, but also account for randomness induced by subgroup identification. We leave this possible extension of the proposed method for future research, as the primary objective of this article is to investigate the causal effect of statin usage on T2D risk in the most vulnerable subgroup.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank the editor, the associate editor, and anonymous reviewers for their comments and suggestions that significantly improved the article. The authors also thank Xuming He and Xinwei Ma for their valuable feedback and thoughtful discussions.

References

Bornkamp B, Ohlssen D, Magnusson BP, and Schmidli H. (2017), "Model Averaging for Treatment Effect Estimation in Subgroups," Pharmaceutical Statistics, 16, 133–142. [1,5] [PubMed: 27935199]

Burke JF, Sussman JB, Kent DM, and Hayward RA (2015), "Three Simple Rules to Ensure Reasonably Credible Subgroup Analyses," Bmj, 351, h5651. [2]

Chen X, and He Y. (2018), "Inference of High-Dimensional Linear Models with Time-Varying Coefficients," Statistica Sinica, 28, 255–276. [10]

Cook D, Brown D, Alexander R, March R, Morgan P, Satterthwaite G, and Pangalos MN (2014), "Lessons Learned from the Fate of Astrazeneca's Drug Pipeline: A Five-Dimensional Framework," Nature Reviews Drug Discovery, 13, 419–431. [5] [PubMed: 24833294]

Dezeure R, Bühlmann P, and Zhang C-H (2017), "High-Dimensional Simultaneous Inference with the Bootstrap," TEST, 26, 685–719. [7,8]

Efron B. (2011), "Tweedie's Formula and Selection Bias," Journal of the American Statistical Association, 106, 1602–1614. [2] [PubMed: 22505788]

Fan A, Song R, and Lu W. (2017), "Change-Plane Analysis for Subgroup Detection and Sample Size Calculation," Journal of the American Statistical Association, 112, 769–778. [2] [PubMed: 28804182]

Friedman J, Hastie T, Simon N, Tibshirani R, Hastie MT, and Matrix D. (2017), "Package 'glmnet.,'" Journal of Statistical Software, 33, 1–22. [6]

Fuentes C, Casella G, and Wells MT (2018), "Confidence Intervals for the Means of the Selected Populations," Electronic Journal of Statistics, 12, 58–79. [2,7,8]

Goodarzi MO, Li X, Krauss RM, Rotter JI, and Chen Y-DI (2013), "Relationship of Sex to Diabetes Risk in Statin Trials," Diabetes Care, 36, e100–e101. [1] [PubMed: 23801803]

Guo X, and He X. (2020), "Inference on Selected Subgroups in Clinical Trials," Journal of the American Statistical Association, 1–18 (just-accepted). [2,5]

Guo X, Wei L, Wu C, and Wang J. (2021), "Sharp Inference on Selected Subgroups in Observational Studies," arXiv preprint arXiv:2102.11338. [5]

Hall P, and Miller H. (2010), "Bootstrap Confidence Intervals and Hypothesis Tests for Extrema of Parameters," Biometrika, 97, 881–892. [2,7]

Holland PW (1986), "Statistics and Causal Inference," Journal of the American statistical Association, 81, 945–960. [2]

Hong L, Kuffner TA, and Martin R. (2018), "On Overfitting and Post-selection Uncertainty Assessments," Biometrika, 105, 221–224. [5]

Ioannidis JP, Tan YJ, and Blum MR (2019), "Limitations and Misinterpretations of e-values for Sensitivity Analyses of Observational Studies," Annals of Internal Medicine, 170, 108–111. [10] [PubMed: 30597486]

Kubota K,Ichinose Y,Scagliotti G,Spigel D,Kim J,Shinkai T,Takeda K, Kim S-W, Hsia T-C, Li R, Tiangco BJ, Yau S, Lim W-T, Yao B, Hei Y-J, and Park K. (2014), "Phase iii Study (MONET1) of Motesanib Plus Carboplatin/Paclitaxel in Patients with Advanced Nonsquamous Nonsmall-Cell Lung Cancer (NSCLC): Asian Subgroup Analysis," Annals of Oncology, 25, 529–536. [1] [PubMed: 24419239]

Lango H, Palmer CN, Morris AD, Zeggini E, Hattersley AT, McCarthy MI, Frayling TM, and Weedon MN (2008), "Assessing the Combined Impact of 18 Common Genetic Variants of Modest Effect Sizes on Type 2 Diabetes Risk," Diabetes, 57, 3129–3135. [3] [PubMed: 18591388]

Li S. (2020), "Debiasing the Debiased Lasso with Bootstrap," Electronic Journal of Statistics, 14, 2298–2337. [5]

Liao KP, Sun J, Cai TA, Link N, Hong C, Huang J, Huffman JE, Gronsbell J, Zhang Y, Ho Y-L, Castro V, Gainer V, Murphy SN, O'Donnell CJ, Michael Gaziano J, Cho K, Szolovits P, Kohane IS, Yu S, and Cai T. (2019), "High-Throughput Multimodal Automated Phenotyping (Map) with Application to PheWAS," Journal of the American Medical Informatics Association, 26, 1255–1262. [4,10] [PubMed: 31613361]

Lipkovich I, Dmitrienko A, Denne J, and Enas G. (2011), "Subgroup Identification based on Differential Effect Search—A Recursive Partitioning Method for Establishing Response to Treatment in Patient Subpopulations," Statistics in Medicine, 30, 2601–2621. [2,4,10] [PubMed: 21786278]

Lu M, Sadiq S, Feaster DJ, and Ishwaran H. (2018), "Estimating Individual Treatment Effect in Observational Data Using Random Forest Methods," Journal of Computational and Graphical Statistics, 27, 209–219. [2] [PubMed: 29706752]

Ma S, and Huang J. (2017), "A Concave Pairwise Fusion Approach to Subgroup Analysis," Journal of the American Statistical Association, 112, 410–423. [2,4,10]

Macedo AF, Douglas I, Smeeth L, Forbes H, and Ebrahim S. (2014), "Statins and the Risk of Type 2 Diabetes Mellitus: Cohort Study Using the UK Clinical Practice Research Datalink," BMC Cardiovascular Disorders, 14, 1–12. [1] [PubMed: 24400643]

Magnusson BP, and Turnbull BW (2013), "Group Sequential Enrichment Design Incorporating Subgroup Selection," Statistics in Medicine, 32, 2695–2714. [2] [PubMed: 23315698]

Mansi I, Frei CR, Wang C-P, and Mortensen EM (2015), "Statins and New-Onset Diabetes Mellitus and Diabetic Complications: A Retrospective Cohort Study of US Halthy Adults," Journal of General Internal Medicine, 30, 1599–1610. [1] [PubMed: 25917657]

Mora S, Glynn RJ, Hsia J, MacFadyen JG, Genest J, and Ridker PM (2010), "Statins for the Primary Prevention of Cardiovascular Events in Women with Elevated High-Sensitivity c-Reactive Protein or Dyslipidemia: Results from the Justification for the Use of Statins in Prevention: An Intervention Trial Evaluating Rosuvastatin (jupiter) and Meta-Analysis of Women from Primary Prevention Trials," Circulation, 121, 1069–1077. [1,2,3,4,10] [PubMed: 20176986]

Mora S, and Ridker PM (2006), "Justification for the Use of Statins in Primary Prevention: An Intervention Trial Evaluating Rosuvastatin (Jupiter)—Can c-Reactive Protein be used to Target Statin Therapy in Primary Prevention?" The American Journal of Cardiology, 97, 33–41. [1]

Nadarajah S, and Kotz S. (2008), "Exact Distribution of the Max/Min of Two Gaussian Random Variables," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 16, 210–212. [5]

Naggara O, Raymond J, Guilbert F, and Altman D. (2011), "The Problem of Subgroup Analyses: An Example from a Trial on Ruptured Intracranial Aneurysms," American Journal of Neuroradiology, 32, 633–636. [1] [PubMed: 21436333]

Nissen SE, Tuzcu EM, Schoenhagen P, Crowe T, Sasiela WJ, Tsai J, Orazem J, Magorien RD, O'Shaughnessy C, and Ganz P. (2005), "Statin Therapy, LDL Cholesterol, c-Reactive Protein, and Coronary Artery Disease," New England Journal of Medicine, 352, 29–38. [1] [PubMed: 15635110]

Packard C, Shepherd J, Cobbe S, Ford I, Isles C, McKillop J, Macfarlane P, Lorimer A, and Norrie J. (1998), "Influence of Pravastatin and Plasma Lipids on Clinical Events in the West of Scotland Coronary Prevention Study (woscops)," Circulation, 97, 1440–1445. [1] [PubMed: 9576423]

Pan W, Sun W, Yang S, Zhuang H, Jiang H, Ju H, Wang D, and Han Y. (2020), "Ldl-c Plays a Causal Role on T2DM: A Mendelian Randomization Analysis," Aging (Albany NY), 12, 2584–2594. [3] [PubMed: 32040442]

Park MY, and Hastie T. (2007), "l_1 Regularization Path Algorithm for Generalized Linear Models," Journal of the Royal Statistical Society, Series B, 69, 659–677. [5]

Povsic TJ, Scott R, Mahaffey KW, Blaustein R, Edelberg JM, Lefkowitz MP, Solomon SD, Fox JC, Healy KE, Khakoo AY, Losordo DW, Malik FI, Monia BP, Montgomery RL, Riesmeyer J, Schwartz GJ, Zelenkofske SL, Wu JC, Wasserman SM, and Roe MT (2017), "Navigating the Future of Ccardiovascular Drug Development—Leveraging Novel Approaches to Drive Innovation and Drug Discovery: Summary of Findings from the Novel Cardiovascular Therapeutics Conference," Cardiovascular Drugs and Therapy, 31, 445–458. [1] [PubMed: 28735360]

Rajpathak SN, Kumbhani DJ, Crandall J, Barzilai N, Alderman M, and Ridker PM (2009), "Statin Therapy and Risk of Developing Type 2 Diabetes: A Meta-Analysis," Diabetes Care, 32, 1924–1929. [1] [PubMed: 19794004]

Robins JM, Rotnitzky A, and Scharfstein DO (2000), "Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models," in Statistical Models in Epidemiology, the Environment, and Clinical Trials, ed. Halloran M. Elizabeth and Donald Berry, pp. 1–94, New York: Springer. [10]

Rosenkranz GK (2016), "Exploratory Subgroup Analysis in Clinical Trials by Model Selection," Biometrical Journal, 58, 1217–1228. [2] [PubMed: 27230820]

Shen J, and He X. (2015), "Inference for Subgroup Analysis with a Structured Logistic-Normal Mixture Model," Journal of the American Statistical Association, 110, 303–312. [2]

Skourtis D, Stavroulaki D, Athanasiou V, Fragouli PG, and Iatrou H. (2020), "Nanostructured Polymeric, Liposomal and Other Materials to Control the Drug Delivery for Cardiovascular Diseases," Pharmaceutics, 12, 1160. [1] [PubMed: 33260547]

Stallard N, Todd S, and Whitehead J. (2008), "Estimation Following Selection of the Largest of Two Normal Means," Journal of Statistical Planning and Inference, 138, 1629–1638. [2]

Sur P, and Candès E. (2019), "A Modern Maximum-Likelihood Theory for High-Dimensional Logistic Regression," Proceedings of the National Academy of Sciences of the United States of America, 116, 14516–14525. [8] [PubMed: 31262828]

Swerdlow DI, Preiss D, Kuchenbaecker KB, Holmes MV, Engmann JE, Shah T, Sofat R, Stender S, Johnson PC, Scott RA, Leusink M, Verweij N, Sharp SJ, Guo Y, Giambartolomei C, Chung C, Peasey A, Amuzu A, Li K, Palmen J, Howard P, Cooper JA, Drenos F, Li YR, Lowe G, Gallacher J, Stewart MC, Tzoulaki I, Buxbaum SG, van der A DL, Forouhi NG, Onland-Moret NC, van der Schouw YT, Schnabel RB, Hubacek JA, Kubinova R, Baceviciene M, Tamosiunas A, Pajak A, Topor-Madry R, Stepaniak U, Malyutina S, Baldassarre D, Sennblad B, Tremoli E, de Faire U, Veglia F, Ford I, Jukema JW, Westendorp RG, de Borst GJ, de Jong PA, Algra A, Spiering W, Maitland-van der Zee AH, Klungel OH, de Boer A, Doevendans PA, Eaton CB, Robinson JG, Duggan D; DIAGRAM Consortium; MAGIC Consortium; InterAct Consortium, Kjekshus J, Downs JR, Gotto AM, Keech AC, Marchioli R, Tognoni G, Sever PS, Poulter NR, Waters DD, Pedersen TR, Amarenco P, Nakamura H, McMurray JJ, Lewsey JD, Chasman DI, Ridker PM, Maggioni AP, Tavazzi L, Ray KK, Seshasai SR, Manson JE, Price JF, Whincup PH, Morris RW, Lawlor DA, Smith GD, Ben-Shlomo Y, Schreiner PJ, Fornage M, Siscovick DS, Cushman M, Kumari M, Wareham NJ, Verschuren WM, Redline S, Patel SR, Whittaker JC, Hamsten A, Delaney JA, Dale C, Gaunt TR, Wong A, Kuh D, Hardy R, Kathiresan S, Castillo BA, van der Harst P, Brunner EJ, Tybjaerg-Hansen A, Marmot MG, Krauss RM, Tsai M, Coresh J, Hoogeveen RC, Psaty BM, Lange LA, Hakonarson H, Dudbridge F, Humphries SE, Talmud PJ, Kivimäki M, Timpson NJ, Langenberg C, Asselbergs FW, Voevoda M, Bobak M, Pikhart H, Wilson JG, Reiner AP, Keating BJ, Hingorani AD, and Sattar N. (2015), "HMG-coenzyme a Reductase Inhibition, Type 2 Diabetes, and Bodyweight: Evidence from Genetic Analysis and Randomised Trials," The Lancet, 385, 351–361. [2,3,4]

Thomas M, and Bornkamp B. (2017), “Comparing Approaches to Treatment Effect Estimation for Subgroups in Clinical Trials,” Statistics in Biopharmaceutical Research, 9, 160–171. [2]

VanderWeele TJ, and Ding P. (2017), “Sensitivity Analysis in Observational Research: Introducing the E-value,” Annals of Internal Medicine, 167, 268–274. [10] [PubMed: 28693043]

Wang J, He X, and Xu G. (2019), “Debiased Inference on Treatment Effect in a High Dimensional Model,” Journal of the American Statistical Association, 1–000 ((just-accepted)). [5]

Wang R, and Ware JH (2013), “Detecting Moderator Effects Using Subgroup Analyses,” Prevention Science, 14, 111–120. [2] [PubMed: 21562742]

Waters DD, Ho JE, Boekholdt SM, DeMicco DA, Kastelein JJ, Messig M, Breazna A, and Pedersen TR (2013), “Cardiovascular Event Reduction Versus New-Onset Diabetes during Atorvastatin Therapy: Effect of Baseline Risk Factors for Diabetes,” Journal of the American College of Cardiology, 61, 148–152. [1,3,4,10] [PubMed: 23219296]

Würtz P, Wang Q, Soininen P, Kangas AJ, Fatemifar G, Tynkkynen T, Tiainen M, Perola M, Tillin T, Hughes AD, Mäntyselkä P, Kähönen M, Lehtimäki T, Sattar N, Hingorani AD, Casas J-P, Salomaa V, Kivimäki M, Järvelin M-R, Smith GD, Vanhala M, Lawlor DA, Raitakari OT, Chaturvedi N, Kettunen J, and AlaKorpela M. (2016), “Metabolomic Profiling of Statin Use and Genetic Inhibition of HMG-CoA Reductase,” Journal of the American College of Cardiology, 67, 1200–1210. [3] [PubMed: 26965542]

Yang S,Lorenzi E,Papadogeorgou G,Wojdyla DM,Li F,andThomas LE (2020), “Propensity Score Weighting for Causal Subgroup Analysis,” arXiv preprint arXiv:2010.02121. [2]

Zhang C-H, and Zhang SS (2014), “Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models,” Journal of the Royal Statistical Society, Series B, 76, 217–242. [5,7,8]

Zöllner S, and Pritchard JK (2007), “Overcoming the Winner’s Curse: Estimating Penetrance Parameters from Case-Control Data,” The American Journal of Human Genetics, 80, 605–615. [5] [PubMed: 17357068]

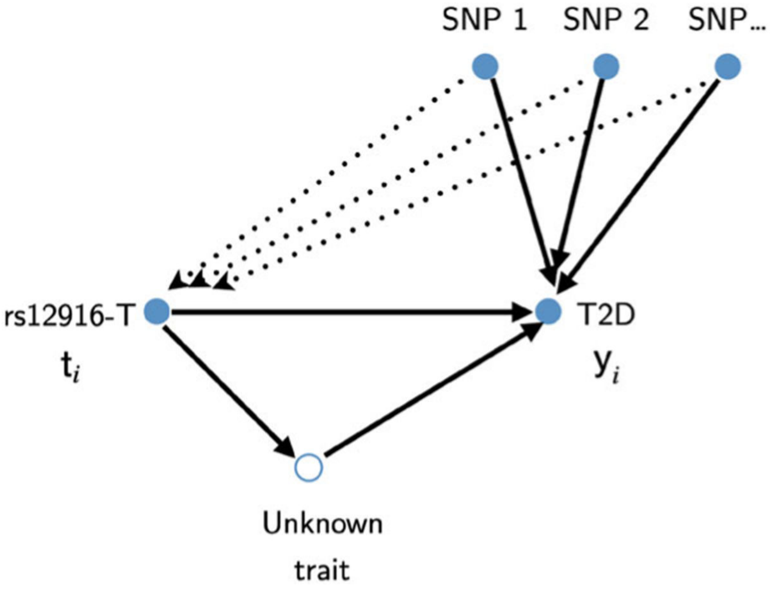

Figure 1.

The causal diagram under our study design.

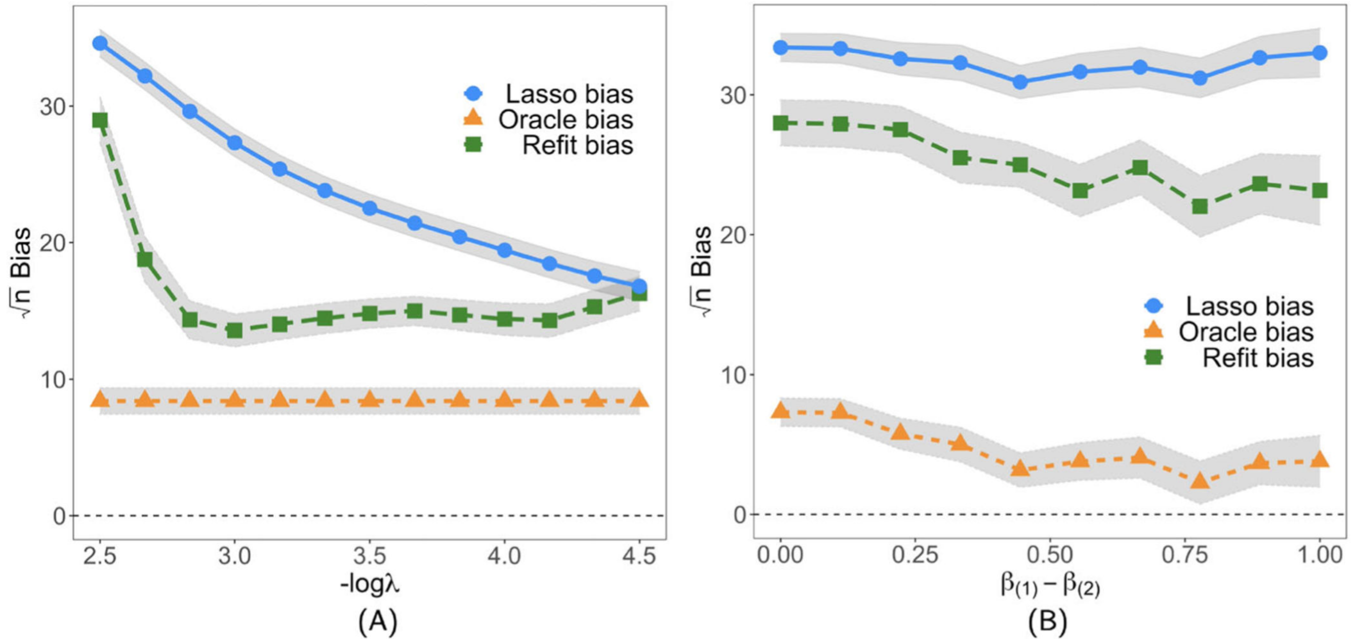

Figure 2.

Root-*n* scaled bias for Example 1. The shaded areas are calculated based on Monte Carlo standard errors.

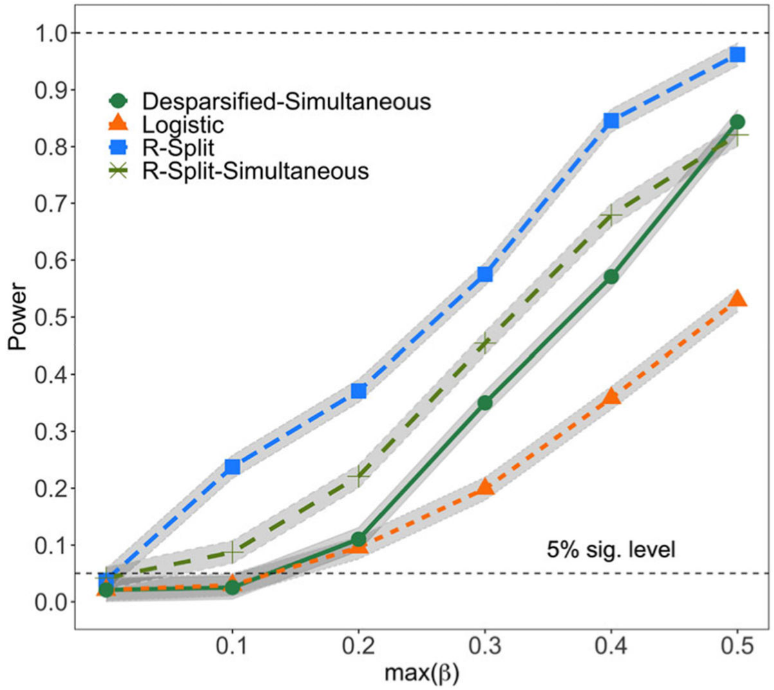

Figure 3.

Power comparison for bootstrap-assisted R-Split, bootstrap-assisted logistic regression, R-Split with simultaneous confidence interval, and the desparsified Lasso with simultaneous confidence interval evaluated over 500 Monte Carlo samples.

Table 1.

Demographics of 17,023 PHS subjects considered in our study.

Variable	Frequency (percent)
Sex	
Female	7592 (45)
Male	9431 (55)
Age (years)	
<40	2333 (14)
40–50	1733 (10)
50–60	2824 (17)
60–70	3971 (23)
70–80	4042 (24)
≥80	2120 (12)
Race	
European	15,048 (88)
African American	1004 (6)
Other/unknown	971 (6)
Ethnicity	
Hispanic or Latino	698 (4)
Other/unknown	16,325 (96)
Number of rs12916-T allele	
2	6245 (37)
1	8079 (47)
0	2699 (16)
With T2D	
Yes	2565 (15)
No	14,458 (85)

Table 2.

The estimated treatment effect (Est), standard error (SE), Z- and (two-sided) p-values of statins' usage on the overall PHS study cohort.

Treatment effect	Est	SE	Z-value	p-value	# of significant coefficients
Full logistic regression	0.04	0.04	0.93	0.35	16

NOTE: We fit the "full" logistic regression model with 337 features (age, race and genetic information). "significant coefficients" are the estimated regression coefficients with p-values < 0.05.

Table 3.

Simulation results (heterogeneous and spurious heterogeneous cases).

		$\beta = (0, \ldots, 0, 1) \in \mathbb{R}^{p_1}$ **(heterogeneity)**			$\beta = (0, \ldots, 0, 0) \in \mathbb{R}^{p_1}$ **(spurious heterogeneity)**		
		Logistic regression ($p_2 = 150$)			**Logistic regression ($p_2 = 150$)**		
		Boot-calibrated	**No adjustment**	**Simultaneous**	**Boot-calibrated**	**No adjustment**	**Simultaneous**
$p_1 = 4$	Cover	0.95(0.01)	0.89(0.01)	0.99(0.01)	0.94(0.02)	0.87(0.03)	0.99(0.01)
	$\sqrt{n}$Length	9.57(0.05)	8.83(0.03)	14.6(0.04)	7.90(0.05)	6.21 (0.05)	11.1 (0.04)
	$\sqrt{n}$Bias	−2.51(2.70)	4.91 (4.60)	—	3.10(3.44)	5.05(4.46)	—
$p_1 = 10$	Cover	0.93(0.01)	0.86(0.01)	0.99(0.01)	0.91(0.01)	0.83(0.02)	0.98(0.01)
	$\sqrt{n}$Length	10.4(0.04)	9.18(0.04)	16.7(0.03)	8.38(0.06)	7.25(0.05)	12.4(0.06)
	$\sqrt{n}$Bias	−4.07(3.89)	5.17(4.67)	—	5.30(4.88)	7.32(6.58)	—
		Repeated sample splitting ($p_2 = 150$)			**Repeated sample splitting ($p_2 = 150$)**		
		Boot-calibrated	**No adjustment**	**Simultaneous**	**Boot-calibrated**	**No adjustment**	**Simultaneous**
$p_1 = 4$	Cover	0.96(0.01)	0.94(0.01)	0.99(0.00)	0.95(0.02)	0.93(0.02)	0.98(0.02)
	$\sqrt{n}$Length	3.56(0.07)	2.17(0.06)	5.14(0.07)	1.87(0.04)	1.03(0.06)	5.08(0.04)
	$\sqrt{n}$Bias	0.11 (0.26)	0.14(0.25)	—	0.15(0.24)	0.31 (0.39)	—
$p_1 = 10$	Cover	0.95(0.02)	0.92(0.01)	0.99(0.01)	0.95(0.01)	0.91 (0.02)	0.96(0.01)
	$\sqrt{n}$Length	3.62(0.07)	2.57(0.05)	6.61 (0.05)	2.02(0.06)	1.47(0.06)	6.46(0.04)
	$\sqrt{n}$Bias	0.25(0.39)	0.32(0.30)	—	0.29(0.40)	0.98(0.90)	—
		Desparsified Lasso ($p_2 = 150$)			**Desparsified Lasso ($p_2 = 150$)**		
		Boot-Calibrated	**No adjustment**	**Simultaneous**	**Boot-Calibrated**	**No adjustment**	**Simultaneous**
$p_1 = 4$	Cover	—	0.92(0.01)	0.99(0.00)	—	0.92(0.01)	0.99(0.01)
	$\sqrt{n}$Length	—	2.13(0.06)	6.51 (0.05)	—	1.01 (0.07)	5.52(0.05)
	$\sqrt{n}$Bias	—	0.29(0.22)	—	—	1.23(0.99)	—
$p_1 = 10$	Cover	—	0.93(0.01)	0.99(0.01)	—	0.93(0.01)	0.99(0.01)

		$\beta = (0, \ldots, 0, 1) \in \mathbb{R}^{p_1}$ **(heterogeneity)**			$\beta = (0, \ldots, 0, 0) \in \mathbb{R}^{p_1}$ **(spurious heterogeneity)**		
		Logistic regression ($p_2 = 150$)			**Logistic regression ($p_2 = 150$)**		
		Boot-calibrated	**No adjustment**	**Simultaneous**	**Boot-calibrated**	**No adjustment**	**Simultaneous**
	$\sqrt{n}$Length	—	2.10(0.07)	6.98(0.07)	—	1.39(0.06)	6.20(0.07)
	$\sqrt{n}$Bias	—	0.27(0.17)	—	—	0.97(0.85)	—
		Repeated sample splitting ($p_2 = 500$)			**Repeated sample splitting ($p_2 = 500$)**		
		Boot-Calibrated	**No adjustment**	**Simultaneous**	**Boot-Calibrated**	**No adjustment**	**Simultaneous**
$p_1 = 4$	Cover	0.95(0.02)	0.92(0.03)	0.99(0.00)	0.95(0.02)	0.91 (0.02)	0.98(0.01)
	$\sqrt{n}$Length	4.44(0.06)	2.22(0.06)	6.08(0.05)	3.77(0.05)	3.18(0.04)	5.90(0.04)
	$\sqrt{n}$Bias	−0.68(0.80)	1.22(1.18)	—	0.62(0.72)	1.58(1.41)	—
$p_1 = 10$	Cover	0.93(0.02)	0.88(0.03)	0.98(0.01)	0.92(0.02)	0.85(0.01)	0.95(0.01)
	$\sqrt{n}$Length	5.11 (0.04)	2.95(0.05)	6.77(0.05)	3.54(0.06)	2.72(0.06)	6.52(0.05)
	$\sqrt{n}$Bias	−0.90(0.85)	1.36(1.20)	—	1.53(1.39)	2.82(1.97)	—
		Desparsified Lasso ($p_2 = 500$)			**Desparsified Lasso ($p_2 = 500$)**		
		Boot-Calibrated	**No adjustment**	**Simultaneous**	**Boot-Calibrated**	**No adjustment**	**Simultaneous**
$p_1 = 4$	Cover	—	0.90(0.01)	0.99(0.00)	—	0.89(0.01)	0.99(0.01)
	$\sqrt{n}$Length	—	2.19(0.05)	7.48(0.08)	—	3.10(0.05)	6.88(0.08)
	$\sqrt{n}$Bias	—	1.29(1.13)	—	—	2.30(1.90)	—
$p_1 = 10$	Cover	—	0.91 (0.01)	0.99(0.01)	—	0.90(0.01)	0.99(0.01)
	$\sqrt{n}$Length	—	2.15(0.06)	7.60(0.08)	—	2.68(0.05)	7.63(0.08)
	$\sqrt{n}$Bias	—	1.25(1.17)	—	—	2.08(1.96)	—

NOTE: “Cover” is the empirical coverage of the 95% lower bound for $\beta_{\max}$. “$\sqrt{n}$Bias” captures the root-*n* scaled Monte Carlo bias for estimating $\beta_{\max}$, and “$\sqrt{n}$ Length” denotes the root-*n* scaled length of the 95% lower bound for $\beta_{\max}$.

Table 4.

Estimated treatment effects (Est) on the PHS cohort in six subgroups divided by sex and T2D genetic risk, together with two-sided 95% confidence intervals (CI), corresponding two-sided *p*-values and the Bonferroni *p*-values in the last column.

Method	Subgroup (prevalence; # of case)	Est (95% CI)	*p*-value	Bonf *p*-value
R-Split (without bootstrap calibration)	High-risk female (0.14,100)	0.41 (0.04, 0.78)	0.030	0.180
	Mid-risk female (0.12,396)	0.10 (−0.03, 0.24)	0.132	0.792
	Low-risk female (0.11,630)	−0.00 (−0.10, 0.09)	0.990	1
	High-risk male (0.24,139)	−0.07 (−0.38, 0.25)	0.658	1
	Mid-risk male (0.21,561)	0.02 (−0.07, 0.11)	0.673	1
	Low-risk male (0.17,739)	−0.03 (−0.16, 0.10)	0.651	1
	Overall	0.07 (−0.16, 0.39)	0.545	-
Simultaneous	High-risk female (0.14,100)	-	0.256	-
Bootstrap-assisted R-Split	High-risk female (0.14,100)	0.35 (0.02, 0.70)	0.037	-

NOTE: We also present the prevalence of T2D in each subgroup.

Table 5.

Sensitivity analysis of our causal evidence measured by the E-value.

Method	Subgroup (prevalence; # of case)	E-value
R-Split (without bootstrap calibration)	High-risk female (0.14, 100)	2.38
	Mid-risk female (0.12, 396)	1.45
	Low-risk female (0.11, 630)	1.00
	High-risk male (0.24, 139)	1.23
	Mid-risk male (0.21, 561)	1.11
	Low-risk male (0.17, 739)	1.14
	Overall	1.23
Bootstrap-assisted R-Split	High-risk female (0.14, 100)	2.19