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

Xinzhou Guo[1*] Waverly Wei[2*] Molei Liu[3] Tianxi Cai[4]
Chong Wu[5] Jingshen Wang[2†]

[1]Department of Mathematics, Hong Kong University of Science and Technology
[2]Division of Biostatistics, UC Berkeley
[3]Department of Biostatistics, Columbia Mailman School of Public Health
[4]Department of Biostatistics, Harvard T.H. Chan School of Public Health
[5]Department of Biostatistics, MD Anderson Cancer Center

Contents

*These authors contributed equally to this work and are alphabetically ordered.
†Correspondence: jingshenwang@berkeley.edu.

A Refitting bias and selection bias

In this section, we demonstrate the refitting bias issues discussed in the main paper with illustrative derivations. To facilitate discussion, suppose for now that $f_i \triangleq \mathrm{expit}'(\mathrm{z}_i^\intercal\beta + \mathrm{x}_i^\intercal\gamma)$ is given. Note that this is infeasible in practice, and neither our theoretical investigation nor our practical implementation requires f_i to be known.

We start with some illustrative derivations on the regularization bias. Since $\beta \in \mathbb{R}^{p_1}$ is a low-dimensional parameter of interest, penalizing β is not necessary in our problem setup. Instead, inference on β can be carried out after a small number of predictors in x are selected (Belloni et al., 2013, 2014). We use $\widehat{\mathtt{M}}$ to record this selected set of predictors. Then, the refitted GLM estimator is obtained via minimizing the negative log-likelihood function

$$(\widehat{\beta}_{\mathrm{GLM}}^\intercal, \widehat{\gamma}_{\mathrm{GLM}}^\intercal)^\intercal = \underset{\beta\in\mathbb{R}^{p_1},\gamma\in\mathbb{R}^{|\widehat{\mathtt{M}}|}}{\arg\min} \left\{ \frac{1}{n}\sum_{i=1}^n \Big(\mathrm{y}_i \cdot (\mathrm{z}_i^\intercal\beta + \mathrm{x}_{i,\widehat{\mathtt{M}}}^\intercal\gamma) - \log\big(1+\exp(\mathrm{z}_i^\intercal\beta + \mathrm{x}_{i,\widehat{\mathtt{M}}}^\intercal\gamma)\big)\Big) \right\}.$$

Under the impact of the random model $\widehat{\mathtt{M}}$ entering the estimation process, $\widehat{\beta}_{\mathrm{GLM}}$ often cannot consistently estimate β unless perfect model selection is achieved (i.e., $\widehat{\mathtt{M}} = \mathtt{M}_0$). To see this, following the derivation provided in the appendix (Section B), we can decompose $\widehat{\beta}_{\mathrm{GLM}}$ into two parts:

$$\sqrt{n}(\widehat{\beta}_{\mathrm{GLM}} - \beta) = \underbrace{\mathrm{I}_{\mathrm{Z}}(\widehat{\Sigma}_{\widehat{\mathtt{M}}})^{-1} \cdot \frac{1}{\sqrt{n}}\sum_{i=1}^n \begin{pmatrix} \mathrm{z}_i \\ \mathrm{x}_{i,\widehat{\mathtt{M}}} \end{pmatrix} \nu_i}_{=:b_{n1}} + \underbrace{(\widetilde{\mathbf{z}}^\intercal(\mathbf{I}-\widetilde{\mathbf{P}}_{\widehat{\mathtt{M}}})\widetilde{\mathbf{z}}/n)^{-1}\widetilde{\mathbf{z}}^\intercal(\mathbf{I}-\widetilde{\mathbf{P}}_{\widehat{\mathtt{M}}})\widetilde{\mathbf{x}}\beta/\sqrt{n}}_{=:b_{n2}} + o_p(1), \tag{1}$$

where $\nu_i = \mathrm{y}_i - \mathrm{expit}(\mathrm{z}_i^\intercal\beta + \mathrm{x}_i^\intercal\gamma)$, I_{Z} denotes an index matrix such that $\mathrm{I}_{\mathrm{Z}}(\mathrm{z}_i^\intercal, \mathrm{x}_{i,\widehat{\mathtt{M}}}^\intercal)^\intercal = \mathrm{z}_i^\intercal$, and $\widehat{\Sigma}_{\widehat{\mathtt{M}}}$ is the sample Hessian matrix defined as $\widehat{\Sigma}_{\widehat{\mathtt{M}}} = \frac{1}{n}\sum_{i=1}^n f_i(\mathrm{z}_i^\intercal, \mathrm{x}_{i,\widehat{\mathtt{M}}}^\intercal)^\intercal(\mathrm{z}_i^\intercal, \mathrm{x}_{i,\widehat{\mathtt{M}}}^\intercal)$. Furthermore, $\mathbf{D} = \mathrm{diag}(\mathbf{D}) = (f_1, \ldots, f_n)$ is a diagonal matrix. Then, $\widetilde{\mathbf{z}}^\intercal = \mathbf{z}^\intercal\mathbf{D}^{1/2}$, $\widetilde{\mathbf{x}}^\intercal = \mathbf{x}^\intercal\mathbf{D}^{1/2}$, $\widetilde{\mathbf{x}}_{\widehat{\mathtt{M}}}^\intercal = \mathbf{x}_{\widehat{\mathtt{M}}}^\intercal\mathbf{D}^{1/2}$, and the projection matrix is $\widetilde{\mathbf{P}}_{\widehat{\mathtt{M}}} = \widetilde{\mathbf{x}}_{\widehat{\mathtt{M}}}(\widetilde{\mathbf{x}}_{\widehat{\mathtt{M}}}^\intercal\widetilde{\mathbf{x}}_{\widehat{\mathtt{M}}})^{-1}\widetilde{\mathbf{x}}_{\widehat{\mathtt{M}}}^\intercal$.

The regularization bias has two sources implied by the decomposition above. The first

bias term b_{n1} is often not centered around zero due to the correlation between ν_i and the data dependent model (i.e., $\mathbb{E}(\nu_i|\mathrm{x}_{i,\widehat{\mathtt{M}}}) \neq 0$). As this bias only occurs whenever an irrelevant variable is selected. The second term b_{n2} captures the impact of omitting variables in the true model $\mathtt{M}_0$ for estimating β, and it occurs whenever the selected model $\widehat{\mathtt{M}}$ under-covers the true support set of γ (i.e., $\mathtt{M}_0$). The impact of the under-fitting vanishes whenever the sure screening property, $\mathtt{M}_0 \subseteq \widehat{\mathtt{M}}$, holds. Existing literature in linear models has argued that sufficient conditions for sure screening property to hold are much weaker than the ones needed for the perfect model selection (Wasserman and Roeder, 2009), indicating selecting a larger model can be a simple remedy to avoid the under-fitting bias.

B Implementation details

Step 1. For $b \leftarrow 1$ to B_1 do

1. Randomly split the data $\{(\mathrm{y}_i, \mathrm{x}_i, \mathrm{z}_i)\}_{i=1}^n$ into group T_1 of size n_1 and group T_2 of size $n_2 = n - n_1$, for $i = 1, \cdots, n$.
2. Select a model $\widehat{\mathtt{M}}_b$ to predict y based on T_1.
3. Refit the model with the data in T_2 to get

$$(\widetilde{\beta}_b^\intercal, \widetilde{\gamma}_b^\intercal)^\intercal = \arg\min \left\{ \sum_{l \in T_2} \left(\mathrm{y}_l \cdot (\mathrm{z}_l^\intercal \beta + \mathrm{x}_{l,\widehat{\mathtt{M}}_b}^\intercal \gamma) - \log\left(1 + \exp(\mathrm{z}_l^\intercal \beta + \mathrm{x}_{l,\widehat{\mathtt{M}}_b}^\intercal \gamma)\right) \right) \right\}.$$

4. Let $f_{bl} = \mathrm{expit}'(\mathrm{z}_l^\intercal \widetilde{\beta}_b + \mathrm{x}_{l,\widehat{\mathtt{M}}_b}^\intercal \widetilde{\gamma}_b)$.
5. The R-Split estimate is obtained by averaging over $\widetilde{\beta}_b$:

$$\widetilde{\beta} = \frac{1}{B_1} \sum_{b=1}^{B_1} \widetilde{\beta}_b.$$

Step 2. 1. For $j \in [p_1]$, calculate:

$$\widetilde{\Gamma}_n = \frac{1}{B_1}\sum_{b=1}^{B_1} \mathrm{I}_z \left(\frac{1}{n_1}\sum_{i=1}^{n} \mathbf{1}_{(i\in T_2)} \cdot f_{bl} \begin{pmatrix} \mathrm{z}_i \\ \mathrm{x}_{i,\widehat{\mathtt{M}}} \end{pmatrix} (\mathrm{z}_i^\intercal, \mathrm{x}_{i,\widehat{\mathtt{M}}}^\intercal)^\intercal \right)^{-1} \mathrm{I}_{\widehat{\mathtt{M}}}, \quad \widetilde{c}_j(r) = (1-n^{r-0.5})(\widetilde{\beta}_{\max}-\widetilde{\beta}_j),$$

where r is a positive tuning parameter between 0 to 0.5.

Step 3. For $b \leftarrow 1$ to B_2 do

1. Generate bootstrap replicate $\widetilde{\beta}^*$:

$$\widetilde{\beta}^* = \widetilde{\beta} + \widetilde{\Gamma}_n \cdot \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} \mathrm{z}_i \\ \mathrm{x}_i \end{pmatrix} \nu_i^*.$$

2. Recalibrate bootstrap statistics via

$$T_b^* = \max_{j\in[p_1]}(\widetilde{\beta}_j^* + \widetilde{c}_j(r)) - \widetilde{\beta}_{\max}.$$

Step 4. The level-α one-sided confidence interval for $\beta_{\max}$ is $[\widetilde{\beta}_{\max} - Q_{T_b^*}(\alpha), +\infty)$, and the level-$\alpha$ two-sided confidence interval for $\beta_{\max}$ is $[\widetilde{\beta}_{\max} - Q_{T_b^*}(\alpha), \widetilde{\beta}_{\max} + Q_{T_b^*}(\alpha)]$ and a bias-reduced estimate.

In Step 1 (1), we recommend a split ratio of $0.6 : 0.4$ for $n_1 : n_2$ because a larger sample size for subsample T_1 improves model selection accuracy. In Step 1 (2), the model selection procedure can be any easily accessible procedure. In our case, since the real data have binary outcomes, we adopt GLM lasso for model selection with R package `glmnet` (Park and Hastie, 2007). The model size is selected via cross-validation with a constraint on the maximal and minimal model sizes. We recommend to set $B_1 = 500$ for the number of repeated splits and $B_2 = 1,000$ for the number of bootstrap replications. As for the tuning parameter r, we propose a data-adaptive cross-validated algorithm to select r as the following (Guo and He, 2020):

Step 1 *Denote* $R = \{r_1, \ldots, r_m\}$ *as a set of candidate tuning parameters. Randomly split the sample into* v *equal-sized subsamples.*

Step 2 *For* $l \leftarrow 1$ *to* m*:*

For $j \leftarrow 1$ *to* v*:*

(a) *Use subsample* j *as reference data and the rest as training data. Obtain* $\widetilde{\beta}_{\max,\text{reduced},j}(r_l)$ *on the training data, where* r_l *is the tuning parameter.*

(b) *For* $i \leftarrow 1$ *to* k*: Obtain R-Split estimate of* $\widetilde{\beta}_{i,j}$ *and its standard error* $\widetilde{\sigma}_{i,j}$ *on the reference data; evaluate the accuracy*

$$h_{i,j}(r_l) = (\widetilde{\beta}_{\max,\text{reduced},j}(r_l) - \widetilde{\beta}_{i,j})^2 - \widetilde{\sigma}^2_{i,j}.$$

Step 3 *Select the tuning parameter via* $\arg\min_{r_l}\{\min_{i\in[k]}[\sum_{j=1}^{j=v} h_{i,j}(r_l)/v]\}$.

Intuitively, we want to choose r that minimizes the mean squared error between the proposed bias reduced estimate $\widetilde{\beta}_{\max,\text{reduced}}$ and $\beta_{\max}$. Because $\beta_{\max}$ is unknown, we provide an approximation of the mean squared error that can be computed via cross-validation in Step 2 (b). The justification of this cross validation method for fixed p_1 can be found in Guo and He (2020). In our empirical work, we implement the above tuning selection method via three-fold cross-validation with a candidate set $R = \{1/3, 1/6, \ldots, 1/30\}$.

C Theoretical investigation for R-Split assisted bootstrap calibration

In this section, first, we show the asymptotic consistency of R-Split estimator under generalized linear models. Second, we show the bootstrap consistency result of R-Split with fixed p_1. Lastly, we provide theoretical details of Theorem 1. Our proofs rely on the following assumptions.

C.1 Assumptions

Assumption 1. *Suppose* $\{(y_i, z_i, x_i)^\intercal\}_{i=1}^n$ *is a random sample and* (z_i, x_i) *have zero mean and bounded support with an upper bound* C*, i.e.* $|z_{ij}| \leq C$*,* $|x_{ij}| \leq C$*,* $|x_{ij}x_{ij}^\intercal| \leq C$*, for* $i = 1, \ldots, n, j = 1, \ldots, p$*.*

Assumption 2. *The split ratio* $r = n_2/n$ *is a constant in* $(0, 1)$*. The selected model sizes in all splits are bounded by* S *with* $S = o(n)$*.*

Assumption 3.

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\left\{\mathbb{E}\Big(s_i a_1^\intercal \widehat{\Sigma}_{\widehat{M},S}^{-1} I_{\widehat{M}}|y,z,x\Big) - \mathbb{E}\Big(\widetilde{s}_i a_1^\intercal \widehat{\Sigma}_{\widetilde{M},\widetilde{S}}^{-1} I_{\widetilde{M},\widetilde{S}}|y,z,x\Big)\right\}(z_i^\intercal, x_i^\intercal)^\intercal \nu_i\right\} = o_p(1),$$

Assumption 4. *For any vector* $a_1 \in \mathbb{R}^{p_1}$*, there exists a random vector* $\eta_n(a_1) \in \mathbb{R}^{p+1}$ *which is independent of* ν*, and* $||\eta(a_1)_n||_\infty$ *is bounded in probability and satisfies*

$$\Big|\Big|r_S\mathbb{E}\Big(a_1^\intercal \widehat{\Sigma}_{\widehat{M},S}^{-1} I_{\widehat{M}}|y,z,x\Big) - \eta_n^\intercal(a_1)\Big|\Big|_1 = o_p(1/\sqrt{\log p}).$$

Assumption 5. *The under-fitting bias over all splits is negligible, such that*

$$\mathbb{E}\Big(\big(\widetilde{\boldsymbol{z}}_S^\intercal(\boldsymbol{I} - \widetilde{\boldsymbol{P}}_{\widehat{M},S})\widetilde{\boldsymbol{z}}_S/n\big)^{-1}\widetilde{\boldsymbol{z}}_S^\intercal(\boldsymbol{I} - \widetilde{\boldsymbol{P}}_{\widehat{M},S})\widetilde{\boldsymbol{x}}_S\beta/\sqrt{n}|y,z,x\Big) = o_p(1).$$

Assumption 6. *We assume* $expit(y|z,x)$ *is continuously differentiable in* y *for each* $(z^\intercal, x^\intercal)$ *in the support of* $(z^\intercal, x^\intercal)$ *and* $|expit'(y|z,x)| \leq C$*, uniformly in* y *and* $(z^\intercal, x^\intercal)$*.*

Assumption 7.

$$\mathbb{E}\Big[\Big(expit(z^\intercal\widehat{\beta} + x^\intercal\widehat{\gamma}) - expit(z^\intercal\beta + x^\intercal\gamma)\Big)^2\Big] = O_p\Big(\frac{|\mathcal{M}_0|\log(p)^{3/2+\delta}}{n}\Big), \quad \delta > 0.$$

Assumption 8. *There exist a constant* U *and* L *where* $L \leq \widetilde{\Sigma}_{n;i,i} \leq U$ *for any* $i \in [p_1]$ *and* $\widetilde{\Sigma}_n^{-1}$ *exists where* $\widetilde{\Sigma}_n$ *is defined in C.3.*

Assumption 9. $\max_{i\in H}\beta_i - \max_{i\notin H}\beta_i \geq \widetilde{L}$ *where* $\widetilde{L}$ *is a constant and* $H = \{j : \beta_j = \beta_{\max}\}$.

Lemma 1. *Under Assumption 1,* $||\boldsymbol{z}^\intercal \nu/\sqrt{n}||_\infty = O_p(\sqrt{\log p})$

Proof. For $K > 0$,

$$
\begin{aligned}
&\mathbb{P}\Big(\max_j \Big|\sum_{i=1}^n \mathrm{z}_{ij}\nu_i/\sqrt{n}\Big| > \sqrt{\log p}K\Big),\\
&\leq \mathbb{E}\Big\{\mathbb{P}\Big(\max_j ||\mathrm{z}_j||_2 \cdot \max_j \Big|\sum_{i=1}^n \frac{\nu_i \mathrm{z}_{ij}}{||\mathrm{z}_j||_2}\Big| > \sqrt{\log p}K\Big|\mathbf{z}\Big)\Big\},\\
&\leq p\mathbb{E}\Big\{\mathbb{P}\Big(\Big|\sum_{i=1}^n \frac{\nu_i \mathrm{z}_{ij}}{||\mathrm{z}_j||_2}\Big| > \sqrt{\log p}K/\max_j ||\mathrm{z}_j||_2\Big|\mathbf{z}\Big)\Big\},\\
&\leq 2\exp\Big(\log p - \frac{\log p K^2}{2\sigma_\nu^2 C^2}\Big),
\end{aligned}
$$

where the last line is by Hoeffding's inequality. □

Assumption 1 applies upper bounds on the covariates. Assumption 2 puts a constraint on the selected model size. Assumption 3 implies that conditioning on subsample S or $\widetilde{S}$ yields the same distributions. Assumption 4 implies $\widehat{\Sigma}_{\widehat{\mathtt{M}}}^{-1}$ converges to a random vector η_n with error rate $1/\sqrt{\log p}$. Assumption 5 assumes the under-fitting bias is negligible (Wang et al., 2019). Assumption 6 assumes smoothness condition for the *expit* function. Assumption 7 provides the convergence rate of GLM Lasso (Farrell, 2015). Assumption 8 basically requires the variance of $\widetilde{\beta}$ is bounded above and below, and Assumption 9 requires that the best subgroup is separable from the second best one.

C.2 Proofs of R-Split's asymptotic normality and bootstrap consistency under GLM

In this section, first, we prove R-Split's asymptotic consistency and normality under GLM as stated in Theorem C.1. Then we show R-Split's bootstrap consistency in the later part of the section.

C.2.1 R-Split's asymptotic and consistency and normality under GLM

Theorem C.1 (Asymptotic normality of R-Split under generalized linear models)**.** *Under Assumptions 1 - 7, the smoothed estimator from R-Split under GLM is asymptotically consistent, such that*

$$\sqrt{n}a_1^\intercal(\widetilde{\beta}-\beta)=\eta_n^\intercal(a_1)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(z_i^\intercal,x_i^\intercal)^\intercal\nu_i+o_p(1),$$

where a_1 is a random vector, $||a_1||_2=1$. $\eta_n(a_1)$ is a random vector as a function of a_1, $\nu_i=y_i-expit(z_i^\intercal\beta+x_i^\intercal\gamma)$. Let $\widetilde{\sigma}=\sigma_\nu\big(\eta_n^\intercal(a_1)\widehat{\Sigma}_n\eta_n(a_1)\big)^{1/2}$,

$$\widetilde{\sigma}^{-1}\sqrt{n}a_1^\intercal(\widetilde{\beta}-\beta)\rightsquigarrow N(0,1),$$

where $\widehat{\Sigma}_n=\frac{1}{n}\sum_{i=1}^{n}f_i(z_i^\intercal,x_i^\intercal)(z_i^\intercal,x_i^\intercal)^\intercal$, $f_i=expit'(z_i^\intercal\beta+x_i^\intercal\gamma)$.

The proof of Theorem C.1 follows three steps: (1) decompose refitting bias, (2) show asymptotic consistency of R-Split under GLM and (3) prove asymptotic normality of R-Split under GLM.

Proof. **Step 1. Refitting bias decomposition**

Here, our goal is to provide refitting bias decomposition under GLM to cast some insights on the refitting bias issue and also simplify the later asymptotic consistency proof. We want to show the refitting bias can be decomposed as

$$\sqrt{n}(\widehat{\beta}_{\mathtt{GLM}}-\beta)=\mathrm{I}_{\mathrm{Z}}(\widehat{\Sigma}_{\widehat{\mathtt{M}}})^{-1}\cdot\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathrm{z}^\intercal,\mathrm{x}_{\widehat{\mathtt{M}}}^\intercal)^\intercal\nu_i+(\widetilde{\mathbf{z}}^\intercal(\mathbf{I}-\widetilde{\mathbf{P}}_{\widehat{\mathtt{M}}})\widetilde{\mathbf{z}}/n)^{-1}\widetilde{\mathbf{z}}^\intercal(\mathbf{I}-\widetilde{\mathbf{P}}_{\widehat{\mathtt{M}}})\widetilde{\mathbf{x}}\beta/\sqrt{n}.\quad(2)$$

To start, since the refitted estimator in GLM satisfies:

$$(\widehat{\beta}_{\mathtt{GLM}}^\intercal,\widehat{\gamma}_{\mathtt{GLM}}^\intercal)^\intercal=\underset{\substack{\beta\in\mathbb{R}^{p_1},\gamma\in\mathbb{R}^{p_2}\\ \gamma_j=0,\ j\notin\widehat{\mathtt{M}}}}{\arg\min}\left\{\frac{1}{n}\sum_{i=1}^{n}\Big(\mathrm{y}_i\cdot(\mathrm{z}_i^\intercal\beta+\mathrm{x}_{i,\widehat{\mathtt{M}}}^\intercal\gamma)-\log\big(1+\exp(\mathrm{z}_i^\intercal\beta+\mathrm{x}_{i,\widehat{\mathtt{M}}}^\intercal\gamma)\big)\Big)\right\},$$

they are the solution to the following equation:

$$\sum_{i=1}^{n} \begin{bmatrix} \mathrm{z}_i \\ \mathrm{x}_{i,\widehat{\mathtt{M}}} \end{bmatrix} \cdot \left[\nu_i + \mathrm{expit}(\mathrm{z}_i^\intercal \beta + \mathrm{x}_i^\intercal \gamma) - \mathrm{expit}(\mathrm{z}_i^\intercal \widehat{\beta}_{\mathtt{GLM}} + \mathrm{x}_i^\intercal \widehat{\gamma}_{\mathtt{GLM}}) \right] = 0,$$

where $\nu_i = \mathrm{y}_i - \mathrm{expit}(\mathrm{z}_i^\intercal \beta + \mathrm{x}_i^\intercal \gamma)$ is a mean-zero random variable. By Taylor expansion, without loss of generality we assume that there exists some intermediate vectors $\widetilde{\beta}_{\mathtt{GLM}} \in (\beta, \widehat{\beta}_{\mathtt{GLM}})$ and $\widetilde{\gamma}_{\mathtt{GLM}} \in (\gamma, \widehat{\gamma}_{\mathtt{GLM}})$ such that

$$\mathrm{expit}(\mathrm{z}_i^\intercal \beta + \mathrm{x}_i^\intercal \gamma) - \mathrm{expit}(\mathrm{z}_i^\intercal \widehat{\beta}_{\mathtt{GLM}} + \mathrm{x}_i^\intercal \widehat{\gamma}_{\mathtt{GLM}}) = \mathrm{expit}'(\mathrm{z}_i^\intercal \widetilde{\beta}_{\mathtt{GLM}} + \mathrm{x}_i^\intercal \widetilde{\gamma}_{\mathtt{GLM}}) \cdot \left[\mathrm{z}_i^\intercal (\beta - \widetilde{\beta}_{\mathtt{GLM}}) + \mathrm{x}_i^\intercal (\gamma - \widetilde{\gamma}_{\mathtt{GLM}}) \right].$$

Thus, by denoting $\mathtt{M}_1 = \{1, \ldots, p_2\} \backslash \widehat{\mathtt{M}}$, we have

$$\begin{aligned}
\sum_{i=1}^{n} \begin{bmatrix} \mathrm{z}_i \\ \mathrm{x}_{i,\widehat{\mathtt{M}}} \end{bmatrix} \cdot \nu_i &= \sum_{i=1}^{n} \mathrm{expit}'(\mathrm{z}_i^\intercal \widetilde{\beta}_{\mathtt{GLM}} + \mathrm{x}_i^\intercal \widetilde{\gamma}_{\mathtt{GLM}}) \begin{bmatrix} \mathrm{z}_i \\ \mathrm{x}_{i,\widehat{\mathtt{M}}} \end{bmatrix} \cdot \left[\mathrm{z}_i^\intercal (\widehat{\beta}_{\mathtt{GLM}} - \beta) + \mathrm{x}_i^\intercal (\widehat{\gamma}_{\mathtt{GLM}} - \gamma) \right], \\
&= \sum_{i=1}^{n} \mathrm{expit}'(\mathrm{z}_i^\intercal \widetilde{\beta}_{\mathtt{GLM}} + \mathrm{x}_i^\intercal \widetilde{\gamma}_{\mathtt{GLM}}) \begin{bmatrix} \mathrm{z}_i \\ \mathrm{x}_{i,\widehat{\mathtt{M}}} \end{bmatrix} \cdot \left[\mathrm{z}_i^\intercal (\widehat{\beta}_{\mathtt{GLM}} - \beta) + \mathrm{x}_{i,\widehat{\mathtt{M}}}^\intercal (\widehat{\gamma}_{\mathtt{GLM},\widehat{\mathtt{M}}} - \gamma_{\widehat{\mathtt{M}}}) + \mathrm{x}_{i,\mathtt{M}_1}^\intercal (\widehat{\gamma}_{\mathtt{GLM},\mathtt{M}_1} - \gamma_{\mathtt{M}_1}) \right], \\
&= \sum_{i=1}^{n} \mathrm{expit}'(\mathrm{z}_i^\intercal \widetilde{\beta}_{\mathtt{GLM}} + \mathrm{x}_i^\intercal \widetilde{\gamma}_{\mathtt{GLM}}) \begin{bmatrix} \mathrm{z}_i \\ \mathrm{x}_{i,\widehat{\mathtt{M}}} \end{bmatrix} \cdot \left[\mathrm{z}_i^\intercal, \mathrm{x}_{i,\widehat{\mathtt{M}}}^\intercal, \mathrm{x}_{i,\mathtt{M}_1}^\intercal \right] \begin{bmatrix} \widehat{\beta}_{\mathtt{GLM}} - \beta \\ \widehat{\gamma}_{\mathtt{GLM},\widehat{\mathtt{M}}} - \gamma_{\widehat{\mathtt{M}}} \\ \widehat{\gamma}_{\mathtt{GLM},\mathtt{M}_1} - \gamma_{\mathtt{M}_1} \end{bmatrix}, \\
&= \sum_{i=1}^{n} \mathrm{expit}'(\mathrm{z}_i^\intercal \widetilde{\beta}_{\mathtt{GLM}} + \mathrm{x}_i^\intercal \widetilde{\gamma}_{\mathtt{GLM}}) \begin{bmatrix} \mathrm{z}_i \\ \mathrm{x}_{i,\widehat{\mathtt{M}}} \end{bmatrix} \cdot \left[\mathrm{z}_i^\intercal, \mathrm{x}_{i,\widehat{\mathtt{M}}}^\intercal \right] \begin{bmatrix} \widehat{\beta}_{\mathtt{GLM}} - \beta \\ \widehat{\gamma}_{\mathtt{GLM},\widehat{\mathtt{M}}} - \gamma_{\widehat{\mathtt{M}}} \end{bmatrix} \\
&\quad + \sum_{i=1}^{n} \mathrm{expit}'(\mathrm{z}_i^\intercal \widetilde{\beta}_{\mathtt{GLM}} + \mathrm{x}_i^\intercal \widetilde{\gamma}_{\mathtt{GLM}}) \begin{bmatrix} \mathrm{z}_i \\ \mathrm{x}_{i,\widehat{\mathtt{M}}} \end{bmatrix} \cdot \mathrm{x}_{i,\mathtt{M}_1}^\intercal \left[\widehat{\gamma}_{\mathtt{GLM},\mathtt{M}_1} - \gamma_{\mathtt{M}_1} \right].
\end{aligned}$$

By Assumption 5, when $\mathtt{M}_0 \subset \widehat{\mathtt{M}}$, the second term is $o_p(1)$. Rearranging the first term, we

have

$$\begin{bmatrix} \widehat{\beta}_{\mathtt{GLM}} - \beta \\ \widehat{\gamma}_{\mathtt{GLM},\widehat{\mathtt{M}}} - \gamma_{\widehat{\mathtt{M}}} \end{bmatrix} = \left(\sum_{i=1}^{n} \mathrm{expit}'(\mathrm{z}_i^\intercal \widetilde{\beta}_{\mathtt{GLM}} + \mathrm{x}_i^\intercal \widetilde{\gamma}_{\mathtt{GLM}}) \begin{bmatrix} \mathrm{z}_i \\ \mathrm{x}_{i,\widehat{\mathtt{M}}} \end{bmatrix} \cdot \left[\mathrm{z}_i^\intercal, \mathrm{x}_{i,\widehat{\mathtt{M}}}^\intercal \right] \right)^{-1} \sum_{i=1}^{n} \begin{bmatrix} \mathrm{z}_i \\ \mathrm{x}_{i,\widehat{\mathtt{M}}} \end{bmatrix} \cdot \nu_i$$
$$- \left(\sum_{i=1}^{n} \mathrm{expit}'(\mathrm{z}_i^\intercal \widetilde{\beta}_{\mathtt{GLM}} + \mathrm{x}_i^\intercal \widetilde{\gamma}_{\mathtt{GLM}}) \begin{bmatrix} \mathrm{z}_i \\ \mathrm{x}_{i,\widehat{\mathtt{M}}} \end{bmatrix} \cdot \left[\mathrm{z}_i^\intercal, \mathrm{x}_{i,\widehat{\mathtt{M}}}^\intercal \right] \right)^{-1}$$
$$\cdot \sum_{i=1}^{n} \mathrm{expit}'(\mathrm{z}_i^\intercal \widetilde{\beta}_{\mathtt{GLM}} + \mathrm{x}_i^\intercal \widetilde{\gamma}_{\mathtt{GLM}}) \begin{bmatrix} \mathrm{z}_i \\ \mathrm{x}_{i,\widehat{\mathtt{M}}} \end{bmatrix} \cdot \mathrm{x}_{i,\mathtt{M}_1}^\intercal \left[\widehat{\gamma}_{\mathtt{GLM},\mathtt{M}_1} - \gamma_{\mathtt{M}_1} \right].$$

Therefore, the refitting bias of GLM estimator $\widehat{\beta}$ can be decomposed as:

$$\sqrt{n}(\widehat{\beta}_{\mathtt{GLM}} - \beta) = \mathrm{I}_{\mathrm{Z}}(\widehat{\Sigma}_{\widehat{\mathtt{M}}})^{-1} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathrm{z}^\intercal, \mathrm{x}_{\widehat{\mathtt{M}}}^\intercal)^\intercal \nu_i + (\widetilde{\mathbf{z}}^\intercal (\mathbf{I} - \widetilde{\mathbf{P}}_{\widehat{\mathtt{M}}}) \widetilde{\mathbf{z}}/n)^{-1} \widetilde{\mathbf{z}}^\intercal (\mathbf{I} - \widetilde{\mathbf{P}}_{\widehat{\mathtt{M}}}) \widetilde{\mathbf{x}} \beta / \sqrt{n},$$

where I_{Z} denotes an index matrix, $\mathrm{I}_{\mathrm{Z}}(\mathrm{z}_i^\intercal, \mathrm{x}_{i,\widehat{\mathtt{M}}}^\intercal)^\intercal = \mathrm{z}_i^\intercal$. Define the sample Hessian matrix as

$$\widehat{\Sigma}_{\widehat{\mathtt{M}}} = \frac{1}{n} \sum_{i=1}^{n} f_i (\mathrm{z}_i^\intercal, \mathrm{x}_{i,\widehat{\mathtt{M}}}^\intercal)(\mathrm{z}_i^\intercal, \mathrm{x}_{i,\widehat{\mathtt{M}}}^\intercal)^\intercal,$$

where $f_i = \mathrm{expit}'(\mathrm{z}_i^\intercal \beta + \mathrm{x}_i^\intercal \gamma)$. Here, we assume f_i is known, for $i = 1, \ldots, n$. Later this section, we relax the strong assumption on f_i and assume f_i is unknown. Denote $\mathbf{D}$ as a diagonal matrix, where $\mathrm{diag}(\mathbf{D}) = (f_1, \ldots, f_n)$. Denote $\widetilde{\mathbf{z}}^\intercal = \mathbf{z}^\intercal \mathbf{D}^{1/2}$, $\widetilde{\mathbf{x}}^\intercal = \mathbf{x}^\intercal \mathbf{D}^{1/2}$ and $\widetilde{\mathbf{x}}_{\widehat{\mathtt{M}}}^\intercal = \mathbf{x}_{\widehat{\mathtt{M}}}^\intercal \mathbf{D}^{1/2}$. Denote the projection matrix as $\widetilde{\mathbf{P}}_{\widehat{\mathtt{M}}} = \widetilde{\mathbf{x}}_{\widehat{\mathtt{M}}} (\widetilde{\mathbf{x}}_{\widehat{\mathtt{M}}}^\intercal \widetilde{\mathbf{x}}_{\widehat{\mathtt{M}}})^{-1} \widetilde{\mathbf{x}}_{\widehat{\mathtt{M}}}^\intercal$.

Step 2. Asymptotic consistency

Now, we want to formally prove the asymptotic consistency of smoothed R-Split estimator under GLM:

$$\sqrt{n} a_1^\intercal (\widetilde{\beta} - \beta) = \eta_n^\intercal(a_1) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathrm{z}_i^\intercal, \mathrm{x}_i^\intercal)^\intercal \nu_i + o_p(1), \tag{3}$$

Step 2 (a). assume f_i is known

In the first part of the proof, for simplicity, we assume f_i is known. (We will assume f_i is unknown and bound the relevant remainder terms in Step 2 (b)). $f_i = \text{expit}'(\mathrm{x}_i^\intercal\beta + \mathrm{z}_i^\intercal\gamma)$ and f_i satisfies Assumption 6. Let $\mathrm{I}_{\widehat{\mathtt{M}}}$ be an index matrix, $\mathrm{I}_{\widehat{\mathtt{M}}}(\mathrm{z}^\intercal, \mathrm{x}^\intercal)^\intercal = (\mathrm{z}^\intercal, \mathrm{x}_{\widehat{\mathtt{M}}}^\intercal)^\intercal$. Take a subsample T_2 of size n_2. Assume the subsample is indexed by $\mathrm{s} = (\mathrm{s}_1, \ldots, \mathrm{s}_n)$, where $\mathrm{s}_i = \mathbf{1}_{(i \in T_2)}$. Denote a_1 as a random vector, where $||a_1||_2 = 1$. For the smoothed estimator $\widetilde{\beta} = \frac{1}{B}\sum_{i=1}^{B} \widehat{\beta}_{\widehat{\mathtt{M}}}$,

$$
\begin{aligned}
\sqrt{n}a_1^\intercal(\widetilde{\beta} - \beta) &= \mathbb{E}\Big(\sqrt{n}(\widehat{\beta}_{\widehat{\mathtt{M}}} - \beta_{\mathtt{M}_0})|\mathrm{y}, \mathrm{z}, \mathrm{x}\Big), \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathbb{E}\Big(a_1^\intercal \widehat{\Sigma}_{\widehat{\mathtt{M}},\mathrm{S}}^{-1} \mathrm{I}_{\widehat{\mathtt{M}}} \cdot \mathrm{s}_i |\mathrm{y}, \mathrm{z}, \mathrm{x}\Big)(\mathrm{z}_i^\intercal, \mathrm{x}_i^\intercal)\nu_i \\
&+ \mathbb{E}\Big(\sqrt{n}(\widetilde{\mathbf{z}}_\mathrm{S}^\intercal(\mathbf{I} - \widetilde{\mathbf{P}}_{\widehat{\mathtt{M}},\mathrm{S}})\widetilde{\mathbf{z}}_\mathrm{S})^{-1}\widetilde{\mathbf{z}}_\mathrm{S}^\intercal(\mathbf{I} - \widetilde{\mathbf{P}}_{\widehat{\mathtt{M}},\mathrm{S}})\widetilde{\mathbf{x}}_\mathrm{S}\beta |\mathrm{y}, \mathrm{z}, \mathrm{x}\Big), \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \eta_n^\intercal(a_1)(\mathrm{z}_i^\intercal, \mathrm{x}_i^\intercal)^\intercal \nu_i + \mathrm{R}_{n1} + \mathrm{R}_{n2},
\end{aligned}
$$

where

$$
\begin{aligned}
\mathrm{R}_{n1} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \Big\{\mathbb{E}\Big(a_1^\intercal \widehat{\Sigma}_{\widehat{\mathtt{M}},\mathrm{S}}^{-1} \mathrm{I}_{\widehat{\mathtt{M}}} \cdot \mathrm{s}_i |\mathrm{y}, \mathrm{z}, \mathrm{x}\Big) - \eta_n^\intercal(a_1)\Big\}(\mathrm{z}_i^\intercal, \mathrm{x}_i^\intercal)^\intercal \nu_i, \\
\mathrm{R}_{n2} &= \mathbb{E}\Big(\sqrt{n}(\widetilde{\mathbf{z}}_\mathrm{S}^\intercal(\mathbf{I} - \widetilde{\mathbf{P}}_{\widehat{\mathtt{M}},\mathrm{S}})\widetilde{\mathbf{z}}_\mathrm{S})^{-1}\widetilde{\mathbf{z}}_\mathrm{S}^\intercal(\mathbf{I} - \widetilde{\mathbf{P}}_{\widehat{\mathtt{M}},\mathrm{S}})\widetilde{\mathbf{x}}_\mathrm{S}\beta |\mathrm{y}, \mathrm{z}, \mathrm{x}\Big).
\end{aligned}
$$

We want to show $\mathrm{R}_{n1} + \mathrm{R}_{n2} = o_p(1)$. Conditioning on $\mathrm{s}_i = 1$,

$$
\begin{aligned}
\mathbb{E}\Big(a_1^\intercal \widehat{\Sigma}_{\widehat{\mathtt{M}},\mathrm{S}}^{-1} \mathrm{I}_{\widehat{\mathtt{M}}} \cdot \mathrm{s}_i |\mathrm{y}, \mathrm{z}, \mathrm{x}\Big) &= \mathbb{E}\Big(a_1^\intercal \widehat{\Sigma}_{\widehat{\mathtt{M}},\mathrm{S}}^{-1} \mathrm{I}_{\widehat{\mathtt{M}}} \cdot \mathrm{s}_i |\mathrm{y}, \mathrm{z}, \mathrm{x}, s_i = 1\Big)\mathbb{P}(\mathrm{s}_i = 1|\mathrm{y}, \mathrm{z}, \mathrm{x}), \\
&= \mathbb{E}\Big(a_1^\intercal \widehat{\Sigma}_{\widehat{\mathtt{M}},\mathrm{S}}^{-1} \mathrm{I}_{\widehat{\mathtt{M}}} |\mathrm{y}, \mathrm{z}, \mathrm{x}, s_i = 1\Big) \cdot r_\mathrm{S}.
\end{aligned}
$$

Assume there is another subsample indexed by $\widetilde{\mathrm{s}} = (\widetilde{\mathrm{s}}_1, \ldots, \widetilde{\mathrm{s}}_n)$, $\widetilde{\mathrm{s}} \perp \mathrm{s}$, where $\widetilde{\mathtt{M}}$ is the selected

model under $\widetilde{\mathrm{s}}$. We can decompose R_{n1} as

$$\mathrm{R}_{n1} = \underbrace{\Big\{\mathrm{r_S}\mathbb{E}\Big(a_1^\intercal \widehat{\Sigma}^{-1}_{\widetilde{M},\widetilde{\mathrm{S}}} \mathrm{I}_{\widetilde{M}} | \mathrm{y},\mathrm{z},\mathrm{x}\Big) - \eta_n^\intercal(a_1)\Big\}^\intercal \frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathrm{z}_i^\intercal, \mathrm{x}_i^\intercal)^\intercal \nu_i}_{\mathrm{R}_{n1,1}},$$
$$+ \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^n \Big\{\mathbb{E}\Big(a_1^\intercal \widehat{\Sigma}^{-1}_{\widehat{\mathtt{M}},\mathrm{S}} \mathrm{I}_{\widehat{\mathtt{M}}} \cdot \mathrm{s}_i | \mathrm{y},\mathrm{z},\mathrm{x}\Big) - \mathbb{E}\Big(a_1^\intercal \widehat{\Sigma}^{-1}_{\widetilde{M},\widetilde{\mathrm{S}}} \mathrm{I}_{\widetilde{M}} \cdot \mathrm{s}_i | \mathrm{y},\mathrm{z},\mathrm{x}\Big)\Big\}^\intercal (\mathrm{z}^\intercal, \mathrm{x}^\intercal)^\intercal \nu_i}_{\mathrm{R}_{n1,2}}.$$

By Assumption 3, $\mathrm{R}_{n1,2} = o_p(1)$. By Hölder's inequality, Assumption 4 and Lemma 1,

$$\mathrm{R}_{n1,1} \leq \Big|\Big|\mathrm{r_S}\mathbb{E}\Big(a_1^\intercal (\widehat{\Sigma}_{\widetilde{\mathtt{M}},\widetilde{\mathrm{S}}})^{-1} \mathrm{I}_{\widetilde{\mathtt{M}}} | \mathrm{y},\mathrm{z},\mathrm{x}\Big) - \eta_n^\intercal(a_1)\Big|\Big|_1 \cdot \Big|\Big|(\mathbf{z}^\intercal, \mathbf{x}^\intercal)^\intercal \nu \sqrt{n}\Big|\Big|_\infty = o_p(1).$$

Therefore, $\mathrm{R}_{n1} + \mathrm{R}_{n2} = o_p(1)$.

Step 2 (b). assume f_i is unknown

Next, we assume f_i is unknown and bound the remainder terms related to f_i. Denote $f_i = \mathrm{expit}'(\mathrm{z}_i^\intercal \beta + \mathrm{x}_i^\intercal \gamma)$ and $\widehat{f}_i = \mathrm{expit}'(\mathrm{z}_i^\intercal \widehat{\beta} + \mathrm{x}_i^\intercal \widehat{\gamma})$. Denote the sample Hessian matrix under true f as $\widehat{\Sigma}_{\widehat{\mathtt{M}},f}$ and under estimated $\widehat{f}$ as $\widehat{\Sigma}_{\widehat{\mathtt{M}},\widehat{f}}$. Similarly, we can define $\widetilde{\mathbf{z}}_f$ and $\widetilde{\mathbf{z}}_{\widehat{f}}$, $\widetilde{\mathbf{x}}_{\widehat{\mathtt{M}},f}$ and $\widetilde{\mathbf{x}}_{\widehat{\mathtt{M}},\widehat{f}}$, $\widetilde{\mathbf{P}}_{\widehat{\mathtt{M}},f}$ and $\widetilde{\mathbf{P}}_{\widehat{\mathtt{M}},\widehat{f}}$. Denote $\nu_i = \big(\mathrm{y}_i - \mathrm{expit}(\mathrm{z}_i^\intercal \beta + \mathrm{x}_i^\intercal \gamma)\big)$ and $\widehat{\nu}_i = \big(\mathrm{y}_i - \mathrm{expit}(\mathrm{z}_i^\intercal \widehat{\beta} + \mathrm{x}_i^\intercal \widehat{\gamma})\big)$.

$$\sqrt{n}(\widehat{\beta}_{\mathtt{GLM}} - \beta) = \mathrm{I_Z} \widehat{\Sigma}^{-1}_{\widehat{\mathtt{M}},f} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathrm{z}_i^\intercal, \mathrm{x}_{i,\widehat{\mathtt{M}}}^\intercal)^\intercal \nu_i + (\widetilde{\mathbf{z}}_{\widehat{f}}^\intercal (\mathbf{I} - \widetilde{\mathbf{P}}_{\widehat{\mathtt{M}},\widehat{f}}) \widetilde{\mathbf{z}}_{\widehat{f}}/n)^{-1} \widetilde{\mathbf{z}}_{\widehat{f}}^\intercal (\mathbf{I} - \widetilde{\mathbf{P}}_{\widehat{\mathtt{M}},\widehat{f}}) \widetilde{\mathbf{x}}_{\widehat{f}} \beta/\sqrt{n} + \mathrm{R}'_n,$$

$$\begin{aligned}
\mathrm{R}'_n &= \mathrm{R}'_{n1} + \mathrm{R}'_{n2}, \\
&= \Big(\mathrm{I_Z} \widehat{\Sigma}^{-1}_{\widehat{\mathtt{M}},\widehat{f}} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathrm{z}_i^\intercal, \mathrm{x}_{i,\widehat{\mathtt{M}}}^\intercal)^\intercal \nu_i - \mathrm{I_Z} \widehat{\Sigma}^{-1}_{\widehat{\mathtt{M}},f} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathrm{z}_i^\intercal, \mathrm{x}_{i,\widehat{\mathtt{M}}}^\intercal)^\intercal \nu_i \Big) \\
&+ \Big(\mathrm{I_Z} \widehat{\Sigma}^{-1}_{\widehat{\mathtt{M}},\widehat{f}} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathrm{z}_i^\intercal, \mathrm{x}_{i,\widehat{\mathtt{M}}}^\intercal)^\intercal \widehat{\nu}_i - \mathrm{I_Z} \widehat{\Sigma}^{-1}_{\widehat{\mathtt{M}},\widehat{f}} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathrm{z}_i^\intercal, \mathrm{x}_{i,\widehat{\mathtt{M}}}^\intercal)^\intercal \nu_i \Big).
\end{aligned}$$

We want to show R'_n is $O_p\Big(\frac{|\mathtt{M}_0|\log(p)^{3/2+\delta}}{n}\Big)$. For simplicity, denote $\mathrm{w}_{i,\widehat{\mathtt{M}}} = (\mathrm{z}_i^\intercal, \mathrm{x}_{i,\widehat{\mathtt{M}}}^\intercal)$.

$$
\begin{aligned}
\mathrm{R}'_{n1} &= a_1^\intercal\Big(\widehat{\Sigma}^{-1}_{\widehat{\mathtt{M}},\widehat{f}} - \widehat{\Sigma}^{-1}_{\widehat{\mathtt{M}},f}\Big)\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathrm{w}^\intercal_{i,\widehat{\mathtt{M}}}\nu_i\Big),\\
&= a_1^\intercal\widehat{\Sigma}^{-1}_{\widehat{\mathtt{M}},\widehat{f}}\Big(\widehat{\Sigma}_{\widehat{\mathtt{M}},f} - \widehat{\Sigma}_{\widehat{\mathtt{M}},\widehat{f}}\Big)\widehat{\Sigma}^{-1}_{\widehat{\mathtt{M}},f}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathrm{w}^\intercal_{i,\widehat{\mathtt{M}}}\nu_i\Big),\\
&= a_1^\intercal\widehat{\Sigma}^{-1}_{\widehat{\mathtt{M}},\widehat{f}}\Big\{\Big(\frac{1}{n}\sum_{i=1}^n f_i\mathrm{w}_{i,\widehat{\mathtt{M}}}\mathrm{w}^\intercal_{i,\widehat{\mathtt{M}}}\Big) - \Big(\frac{1}{n}\sum_{i=1}^n \widehat{f}_i\mathrm{w}_{i,\widehat{\mathtt{M}}}\mathrm{w}^\intercal_{i,\widehat{\mathtt{M}}}\Big)\Big\}\widehat{\Sigma}^{-1}_{\widehat{\mathtt{M}},f}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathrm{w}^\intercal_{i,\widehat{\mathtt{M}}}\nu_i\Big),\\
&\le \max_i\Big|\Big(\frac{1}{n}\sum_{i=1}^n f_i\mathrm{w}_{i,\widehat{\mathtt{M}}}\mathrm{w}^\intercal_{i,\widehat{\mathtt{M}}}\Big) - \Big(\frac{1}{n}\sum_{i=1}^n \widehat{f}_i\mathrm{w}_{i,\widehat{\mathtt{M}}}\mathrm{w}^\intercal_{i,\widehat{\mathtt{M}}}\Big)\Big|\cdot\Big\|a_1^\intercal\widehat{\Sigma}^{-1}_{\widehat{\mathtt{M}},\widehat{f}}\Big\|_2\Big\|\widehat{\Sigma}^{-1}_{\widehat{\mathtt{M}},f}\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathrm{w}^\intercal_{i,\widehat{\mathtt{M}}}\nu_i\Big\|_2.
\end{aligned}
$$

To bound $\max_i\Big|\frac{1}{n}\sum_{i=1}^n(\widehat{f}_i - f_i)\mathrm{w}_{i,\widehat{\mathtt{M}}}\mathrm{w}^\intercal_{i,\widehat{\mathtt{M}}}\Big|$, we first work with $(\widehat{f}_i - f_i)$.

$$
\begin{aligned}
\widehat{f}_i - f_i &= \mathrm{expit}'(\mathrm{z}_i^\intercal\widehat{\beta} + \mathrm{x}_i^\intercal\widehat{\gamma}) - \mathrm{expit}'(\mathrm{z}_i^\intercal\beta + \mathrm{x}_i^\intercal\gamma),\\
&= \mathrm{expit}(\mathrm{z}_i^\intercal\widehat{\beta} + \mathrm{x}_i\widehat{\gamma})\big(1 - \mathrm{expit}(\mathrm{z}_i^\intercal\widehat{\beta} + \mathrm{x}_i\widehat{\gamma})\big) - \mathrm{expit}(\mathrm{z}_i^\intercal\beta + \mathrm{x}_i^\intercal\gamma)(1 - \mathrm{expit}(\mathrm{z}_i^\intercal\beta + \mathrm{x}_i^\intercal\gamma)),\\
&= \Big(\mathrm{expit}(\mathrm{z}_i^\intercal\widehat{\beta} + \mathrm{x}_i\widehat{\gamma}) - \mathrm{expit}(\mathrm{z}_i^\intercal\beta + \mathrm{x}_i^\intercal\gamma)\Big) + \Big(\mathrm{expit}^2(\mathrm{z}_i^\intercal\beta + \mathrm{x}_i\gamma) - \mathrm{expit}^2(\mathrm{z}_i^\intercal\widehat{\beta} + \mathrm{x}_i^\intercal\widehat{\gamma})\Big),\\
&= \Big(\mathrm{expit}(\mathrm{z}_i^\intercal\widehat{\beta} + \mathrm{x}_i\widehat{\gamma}) - \mathrm{expit}(\mathrm{z}_i^\intercal\beta + \mathrm{x}_i^\intercal\gamma)\Big)\\
&+ \Big(\mathrm{expit}(\mathrm{z}_i^\intercal\beta + \mathrm{x}_i\gamma) + \mathrm{expit}(\mathrm{z}_i^\intercal\widehat{\beta} + \mathrm{x}_i^\intercal\widehat{\gamma})\Big)\Big(\mathrm{expit}(\mathrm{z}_i^\intercal\beta + \mathrm{x}_i\gamma) - \mathrm{expit}(\mathrm{z}_i^\intercal\widehat{\beta} + \mathrm{x}_i^\intercal\widehat{\gamma})\Big).
\end{aligned}
$$

Thus $\max_i\Big|\frac{1}{n}\sum_{i=1}^n(\widehat{f}_i - f_i)\mathrm{w}_{i,\widehat{\mathtt{M}}}\mathrm{w}^\intercal_{i,\widehat{\mathtt{M}}}\Big| = C\cdot O_p(\frac{|\mathtt{M}_0|\log(p)^{3/2+\delta}}{n})$ by Assumption 1 and Assumption 7. The last two l_2 norms in R'_{n1} can be bounded by Lemma 1, with convergence rate $O_p(|\widehat{\mathtt{M}}|^{1/2}\sqrt{\log p})$. In sum,

$$
\mathrm{R}'_{n1} = C\cdot O_p(\frac{|\mathtt{M}_0|\log(p)^{3/2+\delta}}{n})\cdot O_p(|\widehat{\mathtt{M}}|^{1/2}\sqrt{\log p}).
$$

Next, we bound the remainder term R'_{n2}.

$$\begin{aligned}
\mathrm{R}'_{n2} &= \Big(a_1^\intercal \widehat{\Sigma}^{-1}_{\widehat{\mathtt{M}},\widehat{f}} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathrm{w}_{i,\widehat{\mathtt{M}}}^\intercal \widehat{\nu}_i - a_1^\intercal \widehat{\Sigma}^{-1}_{\widehat{\mathtt{M}},\widehat{f}} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathrm{w}_{i,\widehat{\mathtt{M}}}^\intercal \nu_i \Big), \\
&= \Big(a_1^\intercal \widehat{\Sigma}^{-1}_{\widehat{\mathtt{M}},\widehat{f}} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathrm{w}_{i,\widehat{\mathtt{M}}}^\intercal (\widehat{\nu}_i - \nu_i) \Big), \\
&= \Big(a_1^\intercal \widehat{\Sigma}^{-1}_{\widehat{\mathtt{M}},\widehat{f}} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathrm{w}_{i,\widehat{\mathtt{M}}}^\intercal \big(\mathrm{expit}(\mathrm{z}_i^\intercal \widehat{\beta} + \mathrm{x}_i^\intercal \widehat{\gamma}) - \mathrm{expit}(\mathrm{z}_i^\intercal \beta + \mathrm{x}_i^\intercal \gamma)\big) \Big), \\
&\leq \sqrt{n} \cdot \lambda_{\min}^{-1}(a_1^\intercal \widehat{\Sigma}_{\widehat{\mathtt{M}},\widehat{f}}) \cdot \sqrt{\mathbb{E}\Big[\big(\mathrm{w}_{i,\widehat{\mathtt{M}}}^\intercal\big)^2\Big] \mathbb{E}\Big[\Big(\mathrm{expit}(\mathrm{z}_i^\intercal \widehat{\beta} + \mathrm{x}_i^\intercal \widehat{\gamma}) - \mathrm{expit}(\mathrm{z}_i^\intercal \beta + \mathrm{x}_i^\intercal \gamma)\Big)^2\Big]}, \\
&= O_p\Big(\frac{|\mathtt{M}_0| \log(p)^{3/2+\delta}}{n}\Big)
\end{aligned}$$

by Hölder's inequality, Assumption 1 and Assumption 7. In sum, $\mathrm{R}'_n = \mathrm{R}'_{n1} + \mathrm{R}'_{n2} = O_p\Big(\frac{|\mathtt{M}_0| \log(p)^{3/2+\delta}}{n}\Big)$. Combining Step 2 (a) and Step 2 (b), we prove the consistency result in Equation (3).

Step 3. Asymptotic normality

Now let $\widetilde{\sigma}_n(a_1) = \sigma_\nu(\eta_n(a_1)^\intercal \widehat{\Sigma}_n \eta_n(a_1))^{1/2}$, where $\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n f_i(\mathrm{z}_i^\intercal, \mathrm{x}_i^\intercal)(\mathrm{z}_i^\intercal, \mathrm{x}_i^\intercal)^\intercal$. Under Assumptions 1, Assumptions 6 and Assumption 7 and the asymptotic consistency result in Equation (3), we have

$$\widetilde{\sigma}_n^{-1}(a_1)\sqrt{n} a_1^\intercal (\widetilde{\beta} - \beta) \rightsquigarrow N(0,1). \tag{4}$$

□

C.2.2 R-Split's bootstrap consistency under GLM

In this section, we want to prove R-Split's bootstrap consistency by showing:

$$a_1^\intercal (\widetilde{\beta}^* - \widetilde{\beta}) = \eta_n^\intercal(a_1) \cdot \frac{1}{n} \sum_{i=1}^{n} (\mathrm{z}_i^\intercal, \mathrm{x}_i^\intercal)^\intercal \nu_i + o_p(1). \tag{5}$$

By the construction of residual bootstrap, we have

$$a_1^\intercal(\widetilde{\beta}^* - \widetilde{\beta}) = \widetilde{\eta}_n^\intercal(a_1) \cdot \frac{1}{n}\sum_{i=1}^n (\mathrm{z}_i^\intercal, \mathrm{x}_i^\intercal)^\intercal \nu_i^*,$$

where $\nu_i^* = u_i\widehat{\nu}_i$. Following a direct expansion of the bootstrap approximation:

$$\widetilde{\eta}_n^\intercal(a_1) \cdot \frac{1}{n}\sum_{i=1}^n (\mathrm{z}_i^\intercal, \mathrm{x}_i^\intercal)^\intercal \nu_i^* = \eta_n^\intercal(a_1) \cdot \frac{1}{n}\sum_{i=1}^n (\mathrm{z}_i^\intercal, \mathrm{x}_i^\intercal)^\intercal u_i\nu_i + r_{n1} + r_{n2},$$

where the remainder terms $r_{n1} = \big(\widetilde{\eta}_n(a_1) - \eta_n(a_1)\big)^\intercal \cdot \frac{1}{n}\sum_{i=1}^n (\mathrm{z}_i^\intercal, \mathrm{x}_i^\intercal)^\intercal u_i\nu_i$ and $r_{n2} = \widetilde{\eta}_n^\intercal(a_1) \cdot \frac{1}{n}\sum_{i=1}^n (\mathrm{z}_i^\intercal, \mathrm{x}_i^\intercal)^\intercal u_i(\widehat{\nu}_i - \nu_i)$. Since u_i's are i.i.d. random variable with mean 0 and variance 1, the leading term has the same distribution as $\widetilde{\beta} - \beta$. Now we define the maximal eigenvalue of matrix $\frac{1}{n}\sum_{i=1}^n (\mathrm{z}_i^\intercal, \mathrm{x}_i^\intercal)^\intercal u_i(\widehat{\nu}_i - \nu_i)$ to be $\lambda_{\max}$, and we assume this quantity is bounded away from infinity. To prove bootstrap consistency, it is suffice to show that the remainder terms vanish at root-n rates. Under the same Assumptions for the previous consistency proof, we immediately have $r_{n1} = o_p(1/\sqrt{n})$. As for the second remainder term, under Assumption 1 and Assumption 7, we have the following bound $r_{n2} \lesssim_p ||\eta_n||_2 \cdot \lambda_{\max} \cdot ||\widehat{\nu} - \nu||_2 = O_p\Big(\frac{|\mathtt{M}_0| \log(p)^{3/2+\delta}}{n}\Big)$.

Following the bootstrap consistency result, we can show $\widetilde{\sigma}_n^{-1}(a_1)\sqrt{n}a_1^\intercal(\widetilde{\beta}^* - \widetilde{\beta}) \rightsquigarrow N(0,1)$ in probability, where $\widetilde{\sigma}_n(a_1) = \sigma_\nu\big(\eta_n^\intercal(a_1)\widehat{\Sigma}_n\eta_n(a_1)\big)^{1/2}$ and $\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n f_i(\mathrm{z}_i^\intercal, \mathrm{x}_i^\intercal)(\mathrm{z}_i^\intercal, \mathrm{x}_i^\intercal)^\intercal$.

C.3 Proof of Theorem 1

By C.1 and C.2, we have under Assumptions 1-7, for any $a \in \mathbb{R}^{p_1}$

$$\widetilde{\sigma}_n^{-1}(a)\sqrt{n}a_1^\intercal(\widetilde{\beta} - \beta) \to N(0,1);\ \ \widetilde{\sigma}_n^{-1}(a)\sqrt{n}a_1^\intercal(\widetilde{\beta}^* - \widetilde{\beta}) \to N(0,1) \textit{ in probability,} \tag{6}$$

Because under Assumption 4, by definition, we can construct a series of $\eta_n(a)$ satisfying additive property; i.e. $\eta_n(k_1c + k_2d) = k_1\eta_n(c) + k_2\eta_n(d)$ for any vectors c and d and constant k_1 and k_2. To be specific, let $\eta_n(e_i)$ denote the random vectors satisfying Assumption 4

for the base vector $e_i \in \mathbb{R}^{p_1}$ for $i = 1, \ldots, p_1$. Then, for any vector $a = \sum_{i=1}^{i=p_1} k_i e_i$, letting $\eta_n(a) = k_i \eta_n(e_i)$ naturally satisfies assumptions 4 as illustrated below,

$$\left|\left|\mathrm{r_S}\mathbb{E}\Big(a^\intercal \widehat{\Sigma}^{-1}_{\widehat{\mathtt{M}},\mathrm{S}} \mathrm{I}_{\widehat{\mathtt{M}}} | \mathrm{y}, \mathrm{z}, \mathrm{x}\Big) - \eta_n^\intercal(a_1)\right|\right|_1 = \left|\left| \sum_{i=1}^{i=p_1} k_i [\mathrm{r_S}\mathbb{E}\Big(e_i^\intercal \widehat{\Sigma}^{-1}_{\widehat{\mathtt{M}},\mathrm{S}} \mathrm{I}_{\widehat{\mathtt{M}}} | \mathrm{y}, \mathrm{z}, \mathrm{x}\Big) - \eta_n^\intercal(e_i)]\right|\right|_1 = o_p(1/\sqrt{\log p}),$$

because by definition, $\left|\left|\mathrm{r_S}\mathbb{E}\Big(e_i^\intercal \widehat{\Sigma}^{-1}_{\widehat{\mathtt{M}},\mathrm{S}} \mathrm{I}_{\widehat{\mathtt{M}}} | \mathrm{y}, \mathrm{z}, \mathrm{x}\Big) - \eta_n^\intercal(e_i)\right|\right|_1 = o_p(1/\sqrt{\log p})$ for $i = 1, \ldots, p_1$. Let $(\xi_{1,n}, \ldots, \xi_{p_1,n}) \sim N(0, \widetilde{\Sigma}_n)$ be a multivariate normal distribution where $\widetilde{\Sigma}_n = \sigma_\nu(L_n \widehat{\Sigma}_n L_n^T)$ and $L_n = (\eta_n(e_1), \ldots, \eta_n(e_{p_1}))^T$. Then, we have $\widetilde{\sigma}_n(a) = a^\intercal \widetilde{\Sigma}_n a$. By Eq. 6 and Assumption 8, because p_1 is fixed, we have

$$\sqrt{n}\widetilde{\Sigma}_n^{-1/2}(\widetilde{\beta} - \beta) \to N(0, I_p); \;\; \sqrt{n}\widetilde{\Sigma}_n^{-1/2}(\widetilde{\beta}^* - \widetilde{\beta}) \to N(0, I_p) \textit{ in probability,} \tag{7}$$

To prove Theorem 1, we prove the following results:

$$\sup_{c \in \mathbb{R}} \left| \mathbb{P}\left(\sqrt{n}\left(\max_{j \in [p_1]} \widetilde{\beta}_j - \beta_{\max}\right) \le c\right) - \mathbb{P}\left(\max_{j \in H} \sqrt{n}(\widetilde{\beta}_j - \beta_j) \le c\right)\right| = o(1),$$
$$\sup_{c \in \mathbb{R}} \left| \mathbb{P}^*\left(\sqrt{n}\left(\max_{j \in [p_1]} (\widetilde{\beta}_j^* + \widetilde{c}_j(r)) - \widetilde{\beta}_{\max}\right) \le c\right) - \mathbb{P}\left(\max_{j \in H} \sqrt{n}(\widetilde{\beta}_j - \beta_j) \le c\right)\right| = o_p(1).$$

First, for the R-Split estimate, we have the following arguments:

$$\begin{aligned}
\sqrt{n}\left(\max_{j \in [p_1]} \widetilde{\beta}_j - \beta_{\max}\right) &= \max_{j \in [p_1]} \left(\sqrt{n}\widetilde{\beta}_j + \sqrt{n}(\beta_j - \beta_j) - \sqrt{n}\beta_{\max}\right) \\
&= \max_{j \in [p_1]} \left(\sqrt{n}(\widetilde{\beta}_j - \beta_j) + \sqrt{n}(\beta_j - \beta_{\max})\right),
\end{aligned}$$

then

$$
\begin{aligned}
&\left|\mathbb{P}\left(\sqrt{n}\left(\max_{j\in[p_1]}\widetilde{\beta}_j-\beta_{\max}\right)\le c\right)-\mathbb{P}\left(\max_{j\in H}\sqrt{n}(\widetilde{\beta}_j-\beta_j)\le c\right)\right| \\
=&\left|\mathbb{P}\left(\max_{j\in[p_1]}\left(\sqrt{n}(\widetilde{\beta}_j-\beta_j)+\sqrt{n}(\beta_j-\beta_{\max})\right)\le c\right)-\mathbb{P}\left(\max_{j\in H}\sqrt{n}(\widetilde{\beta}_j-\beta_j)\le c\right)\right| \\
=&\Big|\mathbb{P}\left(\sqrt{n}(\widetilde{\beta}_j-\beta_j)\le c,\ \text{for } j\in H;\ \sqrt{n}(\widetilde{\beta}_j-\beta_j)+\sqrt{n}(\beta_j-\beta_{\max})\le c,\ \text{for } j\notin H\right) \\
&-\mathbb{P}\left(\max_{j\in H}\sqrt{n}(\widetilde{\beta}_j-\beta_j)\le c\right)\Big| \\
\le&\mathbb{P}\left(\sqrt{n}(\widetilde{\beta}_j-\beta_j)+\sqrt{n}(\beta_j-\beta_{\max})>c,\ \text{any } j\notin H\right) \\
=&1-\mathbb{P}\left(\sqrt{n}(\widetilde{\beta}_j-\beta_j)+\sqrt{n}(\beta_j-\beta_{\max})\le c,\ \text{for } j\notin H\right).
\end{aligned}
$$

Therefore, given any fixed $c_0\in\mathbb{R}$ and for any $c>c_0$, we have

$$
\begin{aligned}
&\left|\mathbb{P}\left(\sqrt{n}\left(\max_{j\in[p_1]}\widetilde{\beta}_j-\beta_{\max}\right)\le c\right)-\mathbb{P}\left(\max_{j\in H}\sqrt{n}(\widetilde{\beta}_j-\beta_j)\le c\right)\right| \\
\le&1-\mathbb{P}\left(\sqrt{n}(\widetilde{\beta}_j-\beta_j)+\sqrt{n}(\beta_j-\beta_{\max})\le c_0,\ \text{for } j\notin H\right).
\end{aligned}
$$

By taking supreme on both side, we obtain

$$
\begin{aligned}
&\sup_{c>c_0}\left|\mathbb{P}\left(\sqrt{n}\left(\max_{j\in[p_1]}\widetilde{\beta}_j-\beta_{\max}\right)\le c\right)-\mathbb{P}\left(\max_{j\in H}\sqrt{n}(\widetilde{\beta}_j-\beta_j)\le c\right)\right| \\
\le&1-\mathbb{P}\left(\sqrt{n}(\widetilde{\beta}_j-\beta_j)\le c_0+\sqrt{n}(\beta_{\max}-\beta_j),\ \text{for } j\notin H\right).
\end{aligned}
$$

By Assumptions 8 and 9 and Eq. 7, the quantity on the right hand side is upper bounded by $\max_{j\in[p_1]}\beta_j-\max_{j\notin H}\beta_j$ and the upper bound of $\widetilde{\Sigma}_{n;i,i}$ over $i\in[p_1]$– the diagnol of $\widetilde{\Sigma}_n$ as

follows

$$
\begin{aligned}
&1-\mathbb{P}\left(\sqrt{n}(\widetilde{\beta}_j-\beta_j)\le c_0+\sqrt{n}(\beta_{\max}-\beta_j),\ \text{for } j\notin H\right)\\
\lesssim&\mathbb{P}\left(\max_{j\notin H}\xi_{j,n}\ge c_0+\sqrt{n}(\max_{j\in[p_1]}\beta_j-\max_{j\notin H}\beta_j)\right)\\
\le&\mathbb{P}\left(\max_{j\in[p_1]}\xi_{j,n}\ge c_0+\sqrt{n}(\max_{j\in[p_1]}\beta_j-\max_{j\notin H}\beta_j)\right)\\
\le&\sum_{i=1}^{i=p_1}\exp\left(-(c_0+\sqrt{n}(\max_{j\in[p_1]}\beta_j-\max_{j\notin H}\beta_j))^2/2\widetilde{\Sigma}_{n;i,i)}\right)\\
\le&\exp\left(\log p_1-\sqrt{n}c_0(\max_{j\in[p_1]}\beta_j-\max_{j\notin H}\beta_j)/U\right)=o(1).
\end{aligned}
$$

where $a_n\lesssim b_n$ indicates that $\limsup_{n\to\infty}a_n\le b_n$.

Since, by Assumption 9, $\limsup_{c_0\to-\infty}\mathbb{P}(\max_{j\in H}\sqrt{n}(\widetilde{\beta}_j-\beta_j)\le c_0)=0$, we have the following statement:

$$
\sup_{c\in\mathbb{R}}\left|\mathbb{P}\left(\sqrt{n}\left(\max_{j\in[p_1]}\widetilde{\beta}_j-\beta_{\max}\right)\le c\right)-\mathbb{P}\left(\max_{j\in H}\sqrt{n}(\widetilde{\beta}_j-\beta_j)\le c\right)\right|=o(1). \tag{8}
$$

Second, since $\widetilde{c}_j(r)=(1-n^{r-0.5})(\widetilde{\beta}_{\max}-\widetilde{\beta}_j)$ the modified bootstrap estimate satisfies

$$
\begin{aligned}
\sqrt{n}\left(\max_{j\in[p_1]}(\widetilde{\beta}_j^*+\widetilde{c}_j(r))-\widetilde{\beta}_{\max}\right)=&\max_{j\in[p_1]}\left(\sqrt{n}\widetilde{\beta}_j^*+(\sqrt{n}-n^r)(\widetilde{\beta}_{\max}-\widetilde{\beta}_j)-\sqrt{n}\widetilde{\beta}_{\max}\right)\\
=&\max_{j\in[p_1]}\left(\sqrt{n}(\widetilde{\beta}_j^*-\widetilde{\beta}_j)+n^r(\widetilde{\beta}_j-\widetilde{\beta}_{\max})\right).
\end{aligned}
$$

Therefore, for $c\in\mathbb{R}$, the distribution of the modified bootstrap estimation has

$$
\begin{aligned}
&\mathbb{P}^*\left(\sqrt{n}\left(\max_{j\in[p_1]}(\widetilde{\beta}_j^*+\widetilde{c}_j(r))-\widetilde{\beta}_{\max}\right)\le c\right)\\
=&\mathbb{P}^*\left(\max_{j\in[p_1]}\left(\sqrt{n}(\widetilde{\beta}_j^*-\widetilde{\beta}_j)+n^r(\widetilde{\beta}_j-\beta_j+\beta_j-\beta_{\max}+\beta_{\max}-\widetilde{\beta}_{\max})\right)\le c\right)\\
=&\mathbb{P}^*\big(\sqrt{n}(\widetilde{\beta}_j^*-\widetilde{\beta}_j)\le c-n^r(\widetilde{\beta}_j-\beta_j+\beta_{\max}-\widetilde{\beta}_{\max}),\ \text{for } j\in H,\\
&\sqrt{n}(\widetilde{\beta}_j^*-\widetilde{\beta}_j)\le c-n^r(\widetilde{\beta}_j-\beta_j+\beta_{\max}-\widetilde{\beta}_{\max})+n^r(\beta_{\max}-\beta_j),\ \text{for } j\notin H\big).
\end{aligned}
$$

Similar to the first part, given any fixed $c_0 \in \mathbb{R}$ and for any $c > c_0$, we have

$$\begin{aligned}
&\sup_{c>c_0} \Big| \mathbb{P}^* \left(\sqrt{n}(\max_{j\in[p_1]}(\widetilde{\beta}_j^* + \widetilde{c}_j(r)) - \widetilde{\beta}_{\max}) \le c \right) \\
&\quad - \mathbb{P}^* \left(\sqrt{n}(\widetilde{\beta}_j^* - \widetilde{\beta}_j) \le c - n^r(\widetilde{\beta}_j - \beta_j + \beta_{\max} - \widetilde{\beta}_{\max}), \text{ for } j \in H \right) \Big| \\
\le& 1 - \mathbb{P}^* \left(\sqrt{n}(\widetilde{\beta}_j^* - \widetilde{\beta}_j) \le c_0 - n^r(\widetilde{\beta}_j - \beta_j + \beta_{\max} - \widetilde{\beta}_{\max}) + n^r(\beta_{\max} - \beta_j), \text{ for } j \notin H \right). \quad (9)
\end{aligned}$$

For the right hand side of (9), recall that under Assumption 8, we have $\max_{j\in[p_1]} n^r|\widetilde{\beta}_j - \beta_j + \beta_{\max} - \widetilde{\beta}_{\max}| = o_p(1)$. By Assumption 8, we have

$$\max_{j\in[p_1]}(n^r|\widetilde{\beta}_j - \beta_j + \beta_{\max} - \widetilde{\beta}_{\max}|/\sqrt{\widehat{\Sigma}_{n;j,j}}) \le \max_{j\in[p_1]}(n^r|\widetilde{\beta}_j - \beta_j + \beta_{\max} - \widetilde{\beta}_{\max}|/\sqrt{L}) = o_p(1)$$

Therefore, by anti-concentration inequality, we have

$$\begin{aligned}
&\mathbb{P}^* \left(\sqrt{n}(\widetilde{\beta}_j^* - \widetilde{\beta}_j) \le c - n^r(\widetilde{\beta}_j - \beta_j + \beta_{\max} - \widetilde{\beta}_{\max}) + n^r(\beta_{\max} - \beta_j), \text{ for } j \notin H \right) \\
&- \mathbb{P} \left(\sqrt{n}(\widetilde{\beta}_j - \beta_j) \le c + n^r(\beta_{\max} - \beta_j), \text{ for } j \notin H \right) = o_p(1),
\end{aligned}$$

uniformly in $c > c_0$. By Assumptions 8 and 9, we can show that

$$1 - \mathbb{P} \left(\sqrt{n}(\widetilde{\beta}_j - \beta_j) \le c + n^r(\beta_{\max} - \beta_j), \text{ for } j \notin H \right) = o(1),$$

uniformly in $c > c_0$. Therefore, the right hand side of (9) converges to 0 in probability. For the left hand side of (9), again by Assumption 8 and anti-concentration inequality, we can also show that

$$\begin{aligned}
&\sup_{c>c_0} \left| \mathbb{P}^* \left(\sqrt{n}(\widetilde{\beta}_j^* - \widetilde{\beta}_j) \le c - n^r(\widetilde{\beta}_j - \beta_j + \beta_{\max} - \widetilde{\beta}_{\max}), \text{ for } j \in H \right) - \mathbb{P} \left(\max_{j\in H} \sqrt{n}(\widetilde{\beta}_j - \beta_j) \le c \right) \right| \\
=& o_p(1).
\end{aligned}$$

By the similar argument we made in the first part, we have shown that

$$\sup_{c\in\mathbb{R}}\left|\mathbb{P}^*\left(\sqrt{n}(\max_{j\in[p_1]}(\widetilde{\beta}_j^*+\widetilde{c}_j(r))-\widetilde{\beta}_{\max})\leq c\right)-\mathbb{P}\left(\max_{j\in H}\sqrt{n}(\widetilde{\beta}_j-\beta_j)\leq c\right)\right|=o_p(1).$$

This result, together with (8) finishes the proof of Theorem 1.

Corollary C.1 (Selected subgroup with the maximal treatment effect)**.** *Under Assumptions 1-9, we have*

$$\sup_{c\in\mathbb{R}}|\mathbb{P}(\sqrt{n}(\widehat{b}_{\max}-\beta_{\widehat{s}})\leq c)-\mathbb{P}^*(\sqrt{n}(\widehat{b}^*_{\text{modified;max}}-\widetilde{\beta}_{\max})\leq c)|=o_p(1).$$

Proof. Let M denote the event $\beta_s<\beta_{\max}$ and s_0 denote the one of the best subgroup; i.e. $\beta_{s_0}=\beta_{\max}$. We have

$$P(M)\leq P(\widetilde{\beta}_{s_0}<\max_{i\notin H}\widetilde{\beta}_i)\leq\sum_{i\notin H}P(\widetilde{\beta}_{s_0}<\widetilde{\beta}_i). \tag{10}$$

Because for any $i\notin H$, by Assumptions 8 and 9, we have

$$P(\widetilde{\beta}_{s_0}<\widetilde{\beta}_i)=P(\sqrt{n}(\widetilde{\beta}_{s_0}-\beta_{s_0}-\widetilde{\beta}_i+\beta_i)<\sqrt{n}(\beta_i-\beta_{s_0}))\to 0$$

as $\sqrt{n}(\beta_i-\beta_{s_0})\to-\infty$. We prove the corollary. □

D Simulation results: Inference on $\beta_{\widehat{S}}$

In this section, we provide additioinal simulatioin results when the inference target is $\beta_{\widehat{s}}$, where $\widehat{s}=\arg\max_{j\in[p_1]}\widehat{\beta}_j$. $\beta_{\widehat{s}}$ denotes the true treatment effect of the selected subgroup $\widehat{s}$. The simulation results are summarized in Table 1.

Table 1: Simulation results (heterogeneous case)

		$\beta = (0, \ldots, 0, 1) \in \mathbb{R}^{p_1}$ (heterogeneity)		
		Logistic Regression ($\boldsymbol{p_2 = 150}$)		
		Boot-Calibrated	No adjustment	Simultaneous
$p_1 = 4$	Cover	0.95(0.02)	0.87(0.02)	0.99(0.01)
	$\sqrt{n}$Length	9.39(0.02)	8.02(0.06)	14.8(0.03)
	$\sqrt{n}$Bias	-3.74(3.37)	5.40(4.09)	—
$p_1 = 10$	Cover	0.92(0.01)	0.85(0.02)	0.98(0.01)
	$\sqrt{n}$Length	10.8(0.05)	9.01(0.02)	16.7(0.02)
	$\sqrt{n}$Bias	-4.77(4.03)	6.06(5.49)	—
		Repeated Sample Splitting ($\boldsymbol{p_2 = 150}$)		
		Boot-Calibrated	No adjustment	Simultaneous
$p_1 = 4$	Cover	0.95(0.01)	0.93(0.02)	0.99(0.01)
	$\sqrt{n}$ Length	3.63(0.07)	2.22(0.05)	5.26(0.06)
	$\sqrt{n}$Bias	0.15(0.32)	0.19(0.27)	—
$p_1 = 10$	Cover	0.94(0.02)	0.92(0.01)	0.99(0.01)
	$\sqrt{n}$Length	3.66(0.05)	2.61(0.05)	6.63(0.06)
	$\sqrt{n}$Bias	0.30(0.42)	0.35(0.32)	—
		Repeated Sample Splitting ($\boldsymbol{p_2 = 500}$)		
		Boot-Calibrated	No adjustment	Simultaneous
$p_1 = 4$	Cover	0.94(0.02)	0.92(0.03)	0.99(0.00)
	$\sqrt{n}$Length	4.45(0.05)	2.25(0.05)	6.10(0.06)
	$\sqrt{n}$Bias	-0.72(0.91)	1.25(1.21)	—
$p_1 = 10$	Cover	0.92(0.03)	0.87(0.02)	0.99(0.00)
	$\sqrt{n}$Length	5.14(0.03)	3.02(0.04)	6.80(0.06)
	$\sqrt{n}$Bias	-1.04(0.89)	1.41(1.22)	—

Note: "Cover" is the empirical coverage of the 95% lower bound for $\beta_{\widehat{s}}$. " $\sqrt{n}$Bias " captures the root-n scaled Monte Carlo bias for estimating $\beta_{\widehat{s}}$, and " $\sqrt{n}$Length " denotes the root-n scaled length of the 95% lower bound for $\beta_{\widehat{s}}$.

E Casual effect identification

E.1 Casual effect identification under the proposed model

In this section, our goal is to showcase that the parameter of interest β indeed represents subgroup treatment effects under Model (11)

$$\text{logit}\left\{\mathbb{P}(\mathrm{y} = 1 \mid \mathrm{z}, \mathrm{x})\right\} = \mathrm{z}^{\intercal}\beta + \mathrm{x}^{\intercal}\gamma, \quad \|\gamma\|_0 \ll p. \tag{11}$$

We work under the Neyman-Rubin (Neyman, 1923; Rubin, 1974) causal model. In accordance with our case study design, each subject is either randomly assigned the treatment, meaning that nature has assigned at least one copy of rs12916-T allele, or the control, mean-

ing that the subject does not inherit rs12916-T allele. The potential outcome y(1) (y(0)) is the potential T2D status we would have observed if the subject carries (does not carry) rs12916-T allele. The observed outcome y = **1**(the subject is diagnosed with T2D) thus equals y = ty(1) + (1 − t)y(0). We work under the stable unit treatment value assumption (SUTVA) and the unconfoundedness assumption listed below.

Assumption 10. *If unit i receives treatment t_i, the observed outcome y_i equals the potential outcome $y_i(t_i)$. In other words, the potential outcome for unit i under treatment t_i is unrelated to the treatment received by other units.*

Assumption 11. *Conditional on a set of potential confounders w, the treatment is independent with the potential outcomes, that is $t \perp\!\!\!\perp y(1), y(0)|w$.*

Since we are interested in the subgroup treatment effect, we use s to denote subgroup indicator variables. We consider six non-overlapping subgroups, $s \in \{1, 2, 3, 4, 5, 6\}$. Because the potential outcomes are binary random variables, we quantify our causal parameter of interest in each subgroup using log odd ratios. For the heterogeneous treatment effect in subgroup $s \in \{1, \ldots, 6\}$, the causal parameter of interest is defined as

$$\begin{aligned}\log \beta_s &= \log \frac{\mathbb{P}\big(\mathrm{y}(1) = 1|\mathrm{s} = s\big)/\big[1 - \mathbb{P}\big(\mathrm{y}(1) = 1|\mathrm{s} = s\big)\big]}{\mathbb{P}(\mathrm{y}(0) = 1|\mathrm{s} = s)/\big[1 - \mathbb{P}\big(\mathrm{y}(0) = 1|\mathrm{s} = s\big)\big]}, \\ &=: \mathrm{logit}\{\mathbb{P}\big(\mathrm{y}(1) = 1|\mathrm{s}\big)\} - \mathrm{logit}\{\mathbb{P}\big(\mathrm{y}(0) = 1|\mathrm{s}\big)\},\end{aligned}$$

The key challenge in causal inference is that for each subject we only observe their potential outcomes under one of the two possible treatments, but never both. Since the potential outcomes are not observed as a priori, our study design aims to enhances the plausibility of the "unconfoundedness assumption" so that causal effects can be identified. The so called unconfoundedness assumption ensures that conditional on a set of potential confounders w, the treatment is independent with the potential outcomes, that is $\mathrm{t} \perp\!\!\!\perp \mathrm{y}(1), \mathrm{y}(0)|\mathrm{w}$.

Our study design enhances the plausibility of the unconfoundedness assumption from two perspectives. On the one hand, the treatment t is a genetic variant, which is randomly

inherited at conception and is not associated with Type 2 diabetes according to GWAS Catalog, therefore it might be reasonable to expect the treatment variable is independent of the potential T2D status acquired after birth (i.e., $\mathrm{t} \perp\!\!\!\perp \mathrm{y}(1), \mathrm{y}(0)$). On the other hand, if one believes that the causal effect between the treatment and the outcome might still be confounded, including prior-birth features (such as age, race, and genetic variant information) as potential confounders makes unconfoundedness assumptions more plausible.

Under the unconfoundedness assumption, we can then identify the causal parameter of interest by conditioning on the confounders w. Take the conditional potential risk in the treated group for subgroup s for example, we identify this causal parameter with

$$\mathbb{P}\big(\mathrm{y}(1) = 1 | \mathrm{s} = s\big) = \mathbb{E}_{\mathrm{W}}\big[\mathbb{P}(\mathrm{y} = 1 | \mathrm{t} = 1, \mathrm{s} = s, \mathrm{w})\big], \quad s \in \{1, \ldots, 6\}.$$

Given the identification condition above, the conditional mean of the outcome model is unknown and therefore needs to be modeled and estimated. In the presence of many potential confounders, we assume that the conditional potential outcome model satisfies

$$\begin{aligned} \operatorname{logit}\big\{\mathbb{P}\big(\mathrm{y}(t) = 1 | \mathrm{w}, \mathrm{s}\big)\big\} &= \operatorname{logit}\big\{\mathbb{P}(\mathrm{y} = 1 | \mathrm{t} = t, \mathrm{s} = s, \mathrm{w})\big\}, & (12) \\ &= \delta_0 + t\delta_1^\intercal s + \delta_2^\intercal s + \delta_3^\intercal \mathrm{w}, & (13) \end{aligned}$$

where w includes age, race, and genetic variants associated with T2D related factors (including LDL, high density lipoprotein and obesity). Model (12) captures our prior belief that the treatment effects can be heterogeneous across the pre-specified subgroups, but may not differ across subpopulations with different genotypes. Model (12) is equivalent to the following logistic regression model with interactions

$$\operatorname{logit}\{\mathbb{P}(\mathrm{y} = 1 | \mathrm{w}, \mathrm{s}, \mathrm{t})\} =: \mathrm{z}^\intercal \beta + \mathrm{x}^\intercal \gamma =: \operatorname{logit}\{\mathbb{P}(\mathrm{y} = 1 | \mathrm{z}, \mathrm{x})\},$$

where $\beta = \big(\log \alpha_1, \ldots, \log \alpha_6\big)$ indeed represents the subgroup treatment effects. $\mathrm{z}^\intercal = \mathrm{ts}^\intercal$,

where $\mathrm{s} = \big(\mathbb{1}(\mathrm{s}_i = 1), \ldots, \mathbb{1}(\mathrm{s}_i = 6)\big)$. $\mathrm{x}^\intercal = (\mathbf{1}^\intercal, \widetilde{\mathrm{s}}^\intercal, \mathrm{w}^\intercal)$, $\widetilde{\mathrm{s}} = \big(\mathbb{1}(\mathrm{s}_i = 1), \ldots, \mathbb{1}(\mathrm{s}_i = 5)\big)$, $\gamma = (\delta_0, \delta_2^\intercal, \delta_3^\intercal)^\intercal$. The derivations are provided in Section E.2 for two-subgroup case and in E.3 for six-subgroup case.

E.2 Parameter identification proof: two subgroups

Proof. Assume an i.i.d. random sample $\{\mathrm{y}_i, \mathrm{t}_i, \mathrm{s}_i, \mathrm{w}_i\}_{i=1}^n$, $\mathrm{s}_i \in \{0, 1\}$. Assume $\mathrm{logit}(\mathbb{P}[\mathrm{y}_i = 1|\mathrm{w}_i, \mathrm{s}_i, \mathrm{t}_i]) = \delta_0 + \delta_1 \mathrm{t}_i \mathrm{s}_i + \delta_2 \mathrm{t}_i(1 - \mathrm{s}_i) + \delta_3 \mathrm{s}_i + \delta_4^\intercal \mathrm{w}_i$.

$$
\begin{aligned}
\mathrm{logit}\Big(\mathbb{P}[\mathrm{y}_i(1) = 1|\mathrm{w}_i, \mathrm{s}_i]\Big) &= \delta_0 + \delta_1 \mathrm{s}_i + \delta_2(1 - \mathrm{s}_i) + \delta_3 \mathrm{s}_i + \delta_4^\intercal \mathrm{w}_i, \\
\mathrm{logit}\Big(\mathbb{P}[\mathrm{y}_i(0) = 1|\mathrm{w}_i, \mathrm{s}_i]\Big) &= \delta_0 + \delta_3 \mathrm{s}_i + \delta_4^\intercal \mathrm{w}_i, \\
\log \alpha_{\mathrm{S}} &= \log\Big(\frac{\mathbb{P}\big(\mathrm{y}(1) = 1|\mathrm{s}_i\big)/\big[1 - \mathbb{P}\big(\mathrm{y}(1) = 1|\mathrm{s}_i\big)\big]}{\mathbb{P}\big(\mathrm{y}(0) = 1|\mathrm{s}_i\big)/\big[1 - \mathbb{P}\big(\mathrm{y}(0) = 1|\mathrm{s}_i\big)\big]}\Big), \\
&= \mathrm{logit}\Big(\mathbb{P}[\mathrm{y}_i(1) = 1|\mathrm{s}_i]\Big) - \mathrm{logit}\Big(\mathbb{P}[\mathrm{y}_i(0) = 1|\mathrm{s}_i]\Big), \\
&= \delta_1 \mathrm{s}_i + \delta_2(1 - \mathrm{s}_i), \\
\mathrm{logit}(\mathbb{P}[\mathrm{y}_i = 1|\mathrm{w}_i, \mathrm{s}_i, \mathrm{t}_i]) &= \mathrm{t}_i \mathrm{logit}\Big(\mathbb{P}[\mathrm{y}_i(1) = 1|\mathrm{w}_i, \mathrm{s}_i]\Big) + (1 - \mathrm{t}_i)\mathrm{logit}\Big(\mathbb{P}[\mathrm{y}_i(0) = 1|\mathrm{w}_i, \mathrm{s}_i]\Big), \\
&= \mathrm{t}_i(\delta_0 + \delta_1 \mathrm{s}_i + \delta_2(1 - \mathrm{s}_i) + \delta_3 \mathrm{s}_i + \delta_4^\intercal \mathrm{w}_i) + (1 - \mathrm{t}_i)(\delta_0 + \delta_3 \mathrm{s}_i + \delta_4^\intercal \mathrm{w}_i), \\
&= \mathrm{t}_i \delta_0 + \mathrm{t}_i \delta_1 \mathrm{s}_i + \mathrm{t}_i \delta_2(1 - \mathrm{s}_i) + \mathrm{t}_i \delta_3 \mathrm{s}_i + \mathrm{t}_i \delta_4^\intercal \mathrm{w}_i \\
&\quad + (1 - \mathrm{t}_i)\delta_0 + (1 - \mathrm{t}_i)\delta_3 \mathrm{s}_i + (1 - \mathrm{t}_i)\delta_4^\intercal \mathrm{w}_i, \\
&= \delta_0 + \mathrm{t}_i \mathrm{s}_i \log \alpha_1 + \mathrm{t}_i(1 - \mathrm{s}_i) \log \alpha_0 + \delta_3 \mathrm{s}_i + \delta_4^\intercal \mathrm{w}_i, \\
\mathrm{logit}(\mathbb{P}[\mathrm{y}_i = 1|\mathrm{w}_i, \mathrm{s}_i, \mathrm{t}_i]) &= \underbrace{\begin{pmatrix} \mathrm{t}_i \mathrm{s}_i & \mathrm{t}_i(1 - \mathrm{s}_i) \end{pmatrix}}_{\mathrm{z}_i^\intercal} \underbrace{\begin{pmatrix} \log \alpha_1 \\ \log \alpha_0 \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} 1 & \mathrm{s}_i & \mathrm{w}_i^\intercal \end{pmatrix}}_{\mathrm{x}_i^\intercal} \underbrace{\begin{pmatrix} \delta_0 \\ \delta_3 \\ \delta_4 \end{pmatrix}}_{\gamma}.
\end{aligned}
$$

The above model is thus equivalent to

$$
\mathrm{logit}\{\mathbb{P}[\mathrm{y}_i = 1|\mathrm{z}_i, \mathrm{x}_i]\} = \mathrm{z}_i^\intercal \beta + \mathrm{x}_i^\intercal \gamma,
$$

where z_i contains the subgroup-treatment interaction terms. $\beta = (\log \alpha_1,\ \log \alpha_0)^\intercal$ represents the subgroup treatment effects. x_i contains an intercept, potential confounders, and subgroup indicator variables. Therefore, the subgroup parameter of interest, β, is identifiable under our proposed model. □

E.3 Parameter identification proof: six subgroups

Proof. Define s_i as a vector of subgroup indicators, $\mathrm{s}_i \in \mathbb{R}^6$. Assume $\text{logit}(\mathbb{P}[\mathrm{y}_i = 1|\mathrm{w}_i, \mathrm{s}_i, \mathrm{t}_i]) = \delta_0 + \mathrm{t}_i\delta_1^\intercal \mathrm{s}_i + \delta_2^\intercal \mathrm{s}_i + \delta_3^\intercal \mathrm{w}_i$.

$$
\begin{aligned}
\text{logit}\Big(\mathbb{P}[\mathrm{y}_i(1) = 1|\mathrm{w}_i, \mathrm{s}_i]\Big) &= \delta_0 + \delta_1^\intercal \mathrm{s}_i + \delta_2^\intercal \mathrm{s}_i + \delta_3^\intercal \mathrm{w}_i, \\
\text{logit}\Big(\mathbb{P}[\mathrm{y}_i(0) = 1|\mathrm{w}_i, \mathrm{s}_i]\Big) &= \delta_0 + \delta_2^\intercal \mathrm{s}_i + \delta_3^\intercal \mathrm{w}_i, \\
\log \alpha_{\mathrm{S}} &= \log\Big(\frac{\mathbb{P}\big(\mathrm{y}(1) = 1|\mathrm{s}_i\big)/\big[1 - \mathbb{P}\big(\mathrm{y}(1) = 1|\mathrm{s}_i\big)\big]}{\mathbb{P}(\mathrm{y}(0) = 1|\mathrm{s}_i)/\big[1 - \mathbb{P}\big(\mathrm{y}(0) = 1|\mathrm{s}_i\big)\big]}\Big), \\
&= \text{logit}\Big(\mathbb{P}[\mathrm{y}_i(1) = 1|\mathrm{s}_i]\Big) - \text{logit}\Big(\mathbb{P}[\mathrm{y}_i(0) = 1|\mathrm{s}_i]\Big), \\
&= \delta_1^\intercal \mathrm{s}_i, \\
\text{logit}(\mathbb{P}[\mathrm{y}_i = 1|\mathrm{w}_i, \mathrm{s}_i, \mathrm{t}_i]) &= \mathrm{t}_i \text{logit}\Big(\mathbb{P}[\mathrm{y}_i(1) = 1|\mathrm{w}_i, \mathrm{s}_i]\Big) + (1 - \mathrm{t}_i)\text{logit}\Big(\mathbb{P}[\mathrm{y}_i(0) = 1|\mathrm{w}_i, \mathrm{s}_i]\Big), \\
&= \mathrm{t}_i(\delta_0 + \delta_1^\intercal \mathrm{s}_i + \delta_2^\intercal \mathrm{s}_i + \delta_3^\intercal \mathrm{w}_i) + (1 - \mathrm{t}_i)(\delta_0 + \delta_2^\intercal \mathrm{s}_i + \delta_3^\intercal \mathrm{w}_i), \\
&= \mathrm{t}_i\delta_0 + \mathrm{t}_i\delta_1^\intercal \mathrm{s}_i + \mathrm{t}_i\delta_2^\intercal \mathrm{s}_i + \mathrm{t}_i\delta_3^\intercal \mathrm{w}_i \\
&\quad + (1 - \mathrm{t}_i)\delta_0 + (1 - \mathrm{t}_i)\delta_2^\intercal \mathrm{s}_i + (1 - \mathrm{t}_i)\delta_3^\intercal \mathrm{w}_i, \\
&= \delta_0 + \mathrm{t}_i \mathrm{s}_i^\intercal \log \alpha_{s_i} + \delta_2^\intercal \mathrm{s}_i + \delta_3^\intercal \mathrm{w}_i, \\
\text{Let } \mathrm{s}_i &= \big(\mathbb{1}(\mathrm{s}_i = 1), \ldots, \mathbb{1}(\mathrm{s}_i = 6)\big)^\intercal,\ \widetilde{\mathrm{s}}_i = \big(\mathbb{1}(\mathrm{s}_i = 1), \ldots, \mathbb{1}(\mathrm{s}_i = 5)\big)^\intercal \\
\log \alpha_{s_i} &= \big(\log \alpha_1, \ldots, \log \alpha_6\big)^\intercal, \\
\text{logit}(\mathbb{P}[\mathrm{y}_i = 1|\mathrm{w}_i, \mathrm{s}_i, \mathrm{t}_i]) &= \underbrace{(\mathrm{t}_i \mathrm{s}_i^\intercal)}_{\mathrm{z}_i^\intercal} \underbrace{(\log \alpha_{s_i})}_{\beta} + \underbrace{(1, \widetilde{\mathrm{s}}_i^\intercal, \mathrm{w}_i^\intercal)}_{\mathrm{x}_i^\intercal} \underbrace{(\delta_0, \delta_3, \delta_4)}_{\gamma}.
\end{aligned}
$$

The above model is thus equivalent to Model (1) considered in the main paper

$$\text{logit}\{\mathbb{P}[\mathrm{y}_i = 1|\mathrm{z}_i, \mathrm{x}_i]\} = \mathrm{z}_i^\intercal \beta + \mathrm{x}_i^\intercal \gamma,$$

where z_i contains the subgroup-treatment interaction terms. $\beta = (\log \alpha_1, \ldots, \log \alpha_6)$ contains the subgroup parameter of interest. x_i contains an intercept, subgroup indicators, and potential confounders. Therefore, the parameter of interest β represents subgroup treatment effect under the proposed model. □

F Additional real data results

F.1 One-sided lower bounds

In this section, we show the real data results with one-sided confidence lower bound. From Table 2, the results of the R-Split estimator without bootstrap calibration suggest that the high-genetic-risk female subgroup is the most vulnerable group for developing T2D with estimated log-odds ratio 0.41, with p-value 0.030, and 95% one-sided confidence lower bound 0.10 (OR = 1.11) with p-value 0.015. Our proposed bootstrap assisted R-Split 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 0.019 for one-sided test).

F.2 Calibration of the second most vulnerable subgroup

As a secondary analysis, we first remove the high-risk female subgroup, which is the most vulnerable subgroup out of the six non-overlapping subgroups. Then the second most vulnerable subgroup (mid-risk female) now becomes the most vulnerable subgroup among the remaining five subgroups. We apply our method to calibrate the estimated treatment effect of the mid-risk female subgroup. The results are summarized in Table 3. Table 3 sug-

Method	Subgroup (prevalence; # of case)	Est (95% LB)	p-value	Bonf p-value
R-Split	High-risk female $(0.14, 100)$	0.41 (0.10)	0.015	0.090
(without bootstrap calibration)	Mid-risk female $(0.12, 396)$	0.10 (-0.01)	0.066	0.396
	Low-risk female $(0.11, 630)$	-0.00 (-0.08)	0.527	1
	High-risk male $(0.24, 139)$	-0.07 (-0.33)	0.664	1
	Mid-risk male $(0.21, 561)$	0.02 (-0.06)	0.336	1
	Low-risk male $(0.17, 739)$	-0.03 (-0.14)	0.667	1
	Overall	0.07 (-0.12)	0.267	–
Simultaneous	High-risk female $(0.14, 100)$	–	0.127	–
Bootstrap-assisted R-Split	High-risk female $(0.14, 100)$	0.35 (0.07)	0.019	–

Table 2: Estimated treatment effects (Est) on the PHS cohort in six subgroups divided by gender and T2D genetic risk, together with the 95% confidence lower bound (LB), the corresponding p-values and the Bonferroni p-values in the last column. We also present the prevalence of T2D in each subgroup.

gests that our method can be naturally applied to correct for the winner's curse bias on the second largest coefficient. After calibration, the mid-risk female subgroup remains to be non-significant.

Method	Subgroup (prevalence; # of case)	Est (95% CI)	p-value	Bonf p-value
R-Split	Mid-risk female $(0.12, 396)$	0.11 $(-0.02,\ 0.23)$	0.10	0.52
(without bootstrap calibration)	Low-risk female $(0.11, 630)$	-0.00 $(-0.16,\ 0.15)$	0.96	1
	High-risk male $(0.24, 139)$	-0.08 $(-0.42,\ 0.26)$	0.64	1
	Mid-risk male $(0.21, 561)$	0.02 $(-0.08,\ 0.13)$	0.67	1
	Low-risk male $(0.17, 739)$	-0.04 $(-0.14,\ 0.06)$	0.50	1
Bootstrap-assisted R-Split	Mid-risk female $(0.12, 396)$	0.04 $(-0.18,\ 0.27)$	0.36	–

Table 3: Estimated treatment effects (Est), their two-sided 95% CI, corresponding two-sided p-values and the Bonferroni p-values after removing the high-risk female subgroup.

G Comparison of pre-defined and post-hoc identified subgroups

In our main paper, "pre-defined subgroups" refers to candidate subgroups that are defined based on prior knowledge, while "post-hoc identified subgroups" refers to the subgroups

identified via data-adaptive identification procedures. Typically, pre-defined subgroups bear better interpretability than post-hoc identified subgroups and avoid the potential bias issue induced by data-adaptively identifying candidate subgroups. Therefore, one considers post-hoc identified subgroups when there is no prior knowledge on subgroup segregation. In our case study, we consider pre-defined subgroups because previous studies (Mora et al., 2010; Waters et al., 2013) suggest that T2D risk might be heterogeneous across sex and T2D genetic profiles.

Although the SNPs we use to define the subgroups are pre-specified using prior knowledge independent of the data, we are able to identify some individualized treatment effects (ITE) and the corresponding subgroups that are similar to our pre-specified subgroups by applying some existing methods to our data. As an illustrative example, we perform logistic regression with Lasso penalty for the model

$$\text{T2D} \sim \text{Age} + \text{Ethnicity} + \text{Treatment} + \text{SNPs} + \text{Treatment} * \text{SNPs}, \tag{14}$$

on all the females. The above model has been widely adopted in subgroup identification literature (Imai et al., 2013; Dusseldorp and Van Mechelen, 2014). We apply Model (14) on females because we aim to investigate if the data-adaptively identified female subgroup also exhibits significant heterogeneous treatment effect as the pre-defined female subgroup in the main paper. Recall that in the main paper, we observe significant treatment effect only in the female high-T2D-risk subgroup. "High-T2D-risk" was defined by the number of risk alleles of rs35011184-A and rs1800961-T.

Here, "SNPs" represents the indicators of having 1 or 2 risk alleles of all the 329 SNPs considered in our case study. Denote the fitted coefficients for the predictors in {Treatment * SNPs} as $\widehat{\zeta}$. We use $\widehat{\zeta}$ to combine and calculate a linear score of SNPs for each individual as the "individualized treatment effect" (ITE), which characterizes the treatment effect heterogeneity of statin usage across different genetic profiles. A higher ITE score represents a higher

genetic risk of developing T2D when treated with statins. From the data, we are able to identify a "high genetic risk" subgroup as those subjects with the top 12% (i.e. prevalence of T2D in the female population) ITE scores based on Model (14). The data-adaptively identified subgroup carries the similar clinical implication to our pre-defined subgroup, that is the female patients who have higher baseline genetic risks of developing T2D.

When comparing the data-adaptively identified subgroup with our pre-defined female high-risk subgroup (i.e. the females with ≥ 2 T2D risk alleles), interestingly, we find that the odds ratio of this data-adaptively identified high-risk group against our pre-defined high-risk subgroup is as high as 2.8, (95% CI: (2.3, 3.4), p-value $< 10^{-16}$). The results demonstrate that the pre-defined subgroup and the data-adaptively identified subgroup are similar.

Please note that identifying subgroups data-adaptively might bring another source of bias. In our main paper, we adopt the subgroups defined based on prior knowledge for analyses which can help avoid such post-selection bias issues. Because correcting the selection bias in data-adaptively identified subgroups is not the main objective of our paper, we shall leave the development of data-adaptive methodologies to future research.

H Discussions on other possible causal pathways

Kindly pointed out by an anonymous referee, there might be patients who did not carry the genetic variant rs12916-T but did take statins to treat diseases such as CAD. In what follows, we shall demonstrate that although there exist patients who did not carry this variant but did take statins, our current study design still provides valid causal effect estimates.

Given that our study cohort may contain patients who did not carry the variant rs12916-T but did take statins, including these patients in our causal analysis potentially opens two different causal pathways including $\mathrm{t}_i \rightarrow \mathrm{w}_i \leftarrow \mathrm{y}_i$ and $\mathrm{t}_i \rightarrow \mathrm{w}_i \rightarrow \mathrm{y}_i$, where w_i represents statin use information after birth. We reflect these additional pathways in the causal diagram in Figure 1. In the causal diagram provided in Figure 1, there are three causal pathways

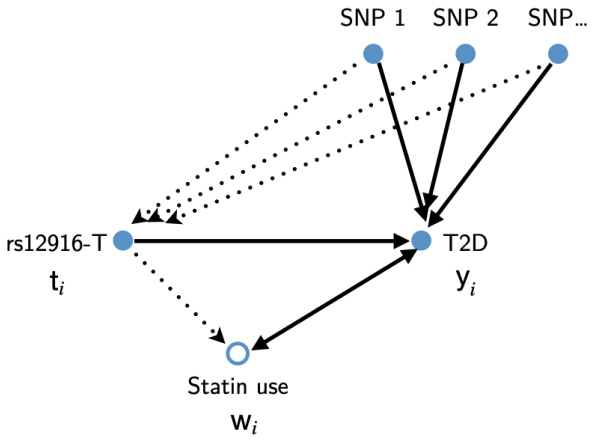

Figure 1: The causal diagram of incorporating statin use information. "Statin use" is a mediator in this causal diagram

started from t_i and ended in y_i, including $t_i \to w_i \to y_i$, $t_i \to w_i \leftarrow y_i$, and $t_i \leftarrow SNP_j \to y_i$, where w_i represents statin use information after birth. In what follows, we shall analyze each pathway and demonstrate that only our current causal pathway $t_i \leftarrow SNP_j \to y_i$ in this case study provide valid causal conclusion.

By the D-separation criteria (Pearl et al. (2016), Ch.3), there are two creditable causal pathways ($t_i \to w_i \to y_i$ and $t_i \leftarrow SNP_j \to y_i$) present in Figure 1, meaning that conducting causal analysis with either of them would lead to valid causal conclusions. However, only the pathway $t_i \leftarrow SNP_j \to y_i$ allows us to establish valid causal conclusion in our study cohort.

For the pathway $t_i \to w_i \to y_i$, under appropriate conditions provided in Imai et al. (2010), this pathway indeed allows us to estimate the direct effect of inheriting rs12916-t on T2D risk. Unfortunately, because statin use information collected after birth is not available in our EHR data, we were not able to conduct causal inference following this pathway.

For the pathway $t_i \leftarrow SNP_j \to y_i$, not only all confounder information is available to us in the EHR data, our study design also guarantees that the cause (carrying rs12916-T [proxy for pharmacological action of statin use] or not) must occur before T2D onset as genetic variants are randomly inherited at conception. Therefore, we are able to establish a clear causal direction between the treatment and the outcome. On top of valid causal directions, following the causal parameter identification proof in Supplementary Materials Section E,

we guarantee that we are able to estimate the causal effect from this causal pathway.

For the pathway $t_i \rightarrow w_i \leftarrow y_i$, because w_i is a collider, including this variable in the causal analysis will create a non-causal association between t_i and y_i. This pathway thus should not be considered in our analysis.

Furthermore, we want to note that even if statin use information is available, incorporating statin use information under our electronic health records (EHR) data is infeasible. This is because the temporal precedence between statin use and T2D onset is not available, the causal direction between t_i and w_i can not be established. In sum, we are not able to decide if statin use is a collider or a mediator between t_i and y_i.

Lastly, while rs12916-T is sometimes used as an instrumental variable in Mendelian Randomization analyses (Würtz et al., 2016), we use rs12916-T as a surrogate treatment variable, instead of an instrumental variable. Our study design is thus different from Mendelian randomization (MR) (Kang et al., 2016; Windmeijer et al., 2019). Furthermore, because MR often assumes that the causal direction between two traits is known as a priori (Xue and Pan, 2020) and our data neither contain statin use information after birth nor provide the temporal order between statin usage and T2D onset, the causal direction for MR analyses can not be specified and MR is not suitable in our case study.

I Sensitivity analysis on the surrogate outcome

Given that our outcome is an error-prone surrogate of the true disease status, we conduct a sensitivity analysis regarding the potential misspecification of the logistic regression model for the true EHR disease status against the covariates. This sensitivity analysis is inspired by Hong et al. (2019) and Zhang et al. (2020). Denote y_i as the observed EHR surrogate of T2D disease status, y_i^* as the unobserved true disease status of T2D. Following Hong et al. (2019), we assume that

$$\text{logit}\left\{\mathbb{P}(y^* = 1 \mid z, x)\right\} = z^\intercal \beta + x^\intercal \gamma, \quad y \perp (z, x) \mid y^*. \tag{15}$$

Eq (15) entails two model assumptions. First, the true disease status y^* follows a logistic regression model against the subgroup-treatment interaction terms z and baseline covariates x (e.g. genetic and demographic variables). Second, by conditioning on y^*, the surrogate outcome y is independent of the predictors, that is the surrogate outcome obtained from phenotyping algorithms is only related to the baseline covariates x through y^*. Under these model assumptions, the log-likelihood function for $\{(\mathrm{y}_i, \mathrm{x}_i, \mathrm{z}_i)\}_{i=1}^n$ can be written as

$$\begin{aligned}\mathcal{L}(\beta, \gamma, \mu) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{y}_i \log\left\{\mu_1 g(\mathrm{z}_i^\intercal \beta + \mathrm{x}_i^\intercal \gamma) + \mu_0 \bar{g}(\mathrm{z}_i^\intercal \beta + \mathrm{x}_i^\intercal \gamma)\right\} \\ + (1-\mathrm{y}_i)\log\left\{(1-\mu_1) g(\mathrm{z}_i^\intercal \beta + \mathrm{x}_i^\intercal \gamma) + (1-\mu_0)\bar{g}(\mathrm{z}_i^\intercal \beta + \mathrm{x}_i^\intercal \gamma)\right\},\end{aligned} \tag{16}$$

where $\mu = (\mu_0, \mu_1)$, $\mu_0 = \mathbb{P}(\mathrm{y} = 1 \mid \mathrm{y}^* = 0)$, $\mu_1 = \mathbb{P}(\mathrm{y} = 1 \mid \mathrm{y}^* = 1)$, $g(\cdot) = \mathrm{logit}^{-1}(\cdot)$, and $\bar{g}(\cdot) = 1 - g(\cdot)$. Due to the high dimensionality of γ, we introduce Lasso penalty and adopte an EM algorithm to solve

$$\{\widehat{\beta}, \widehat{\gamma}, \widehat{\mu}\} = \arg\min_{(\theta,\beta)} \left\{-\mathcal{L}(\beta, \gamma, \mu) + \lambda\|\gamma\|_1\right\},$$

where λ is the penalty parameter. Details of the EM algorithm can be found in Hong et al. (2019). Then we can derive the conditional mean of the true T2D status y_i^* given $(\mathrm{y}_i, \mathrm{x}_i, \mathrm{z}_i)$ by

$$\widehat{\mathrm{y}}_i^* = \frac{\mu_1^{\mathrm{y}_i}(1-\mu_1)^{1-\mathrm{y}_i} g(\mathrm{z}_i^\intercal \widehat{\beta} + \mathrm{x}_i^\intercal \widehat{\gamma})}{\mu_1^{\mathrm{y}_i}(1-\mu_1)^{1-\mathrm{y}_i} g(\mathrm{z}_i^\intercal \widehat{\beta} + \mathrm{x}_i^\intercal \widehat{\gamma}) + \mu_0^{\mathrm{y}_i}(1-\mu_0)^{1-\mathrm{y}_i} \bar{g}(\mathrm{z}_i^\intercal \widehat{\beta} + \mathrm{x}_i^\intercal \widehat{\gamma})}.$$

Finally, for calibration of the error-prone surrogate outcome y_i, we sample the T2D status $\widetilde{\mathrm{y}}_i^*$ following $\mathbb{P}(\widetilde{\mathrm{y}}_i^* = 1 \mid \mathrm{y}_i, \mathrm{z}_i, \mathrm{x}_i) = \widehat{\mathrm{y}}_i^*$, for $i = 1, 2, \ldots, n$, and implement the bootstrap-assisted R-split with $\widetilde{\mathrm{y}}_i^*$ against $(\mathrm{z}_i, \mathrm{x}_i)$. Because this sensitivity analysis uses the calibrated $\widetilde{\mathrm{y}}_i^*$ instead of y_i as the outcome, the sensitivity analysis can correct for the approximation error of the EHR outcome y_i to the true disease status y_i^* (Hong et al., 2019).

To avoid over-fitting bias induced by estimating parameters $(\{\widehat{\beta}, \widehat{\gamma}, \widehat{\mu}\})$ and constructing $\widehat{\mathrm{y}}_i^*$ on the same data, we use a cross-fitting strategy that splits the data into five folds,

estimates $\{\widehat{\beta}, \widehat{\gamma}, \widehat{\mu}\}$ leaving out one fold each time, and constructs each $\widehat{\mathrm{y}}_i^*$ on the left-out fold with the independent estimators. We replicate the sampling of $\{\widetilde{\mathrm{y}}_i^* : i = 1, 2, \ldots, n\}$ for 10 times and take the average over estimated $\beta_{\max}$'s and the associated standard errors under $(\widetilde{\mathrm{y}}_i^*, \mathrm{z}_i, \mathrm{x}_i)$. The resulted point estimates, confidence intervals, and p-values are presented in Table 4.

Method	Subgroup (prevalence; # of case)	Est (95% CI)	p-value	Bonf p-value
R-Split	High-risk female $(0.14, 100)$	0.36 (0.05, 0.67)	0.024	0.144
(without bootstrap calibration)	Mid-risk female $(0.12, 396)$	0.09 (-0.07, 0.25)	0.275	1
	Low-risk female $(0.11, 630)$	0.03 (-0.09, 0.15)	0.651	1
	High-risk male $(0.24, 139)$	-0.05 (-0.36, 0.26)	0.754	1
	Mid-risk male $(0.21, 561)$	-0.01 (-0.14, 0.13)	0.940	1
	Low-risk male $(0.17, 739)$	-0.01 (-0.13, 0.11)	0.886	1
Bootstrap-assisted R-Split	High-risk female $(0.14, 100)$	0.28 (0.02, 0.54)	0.037	–

Table 4: The sensitivity analysis of our surrogate outcome, including the estimated treatment effects (Est), their two-sided 95% confidence intervals (CI), the two-sided p-values, and the Bonferroni p-values obtained by implementing the R-Split and the Bootstrap-assisted R-split procedures with the calibrated outcome $\widetilde{\mathrm{y}}_i^*$ (instead of y_i) against $(\mathrm{z}_i, \mathrm{x}_i)$. The results are produced by averaging over the results from 10 repetitions of sampling $\widetilde{\mathrm{y}}_i^*$.

Comparing the results in Table 4 with Table 4 in the main paper, we do not observe any significant differences. In both tables, R-split p-values of the high-risk female group are around 0.03, the p-values of the remaining subgroups are non-significant, and the bootstrap-assisted R-Split p-values are equal to 0.037, which lead to the same scientific conclusion as in Table 4 in the main paper. The results from the sensitivity analysis suggest that the analyses and findings in Table 4 in the main paper are not sensitive to the approximation error of the EHR surrogate y_i to the true T2D status. This is because our EHR outcome y, derived using MAP, shows a very low approximation error to the true T2D status (AUC = 0.99, specificity = 0.97, and sensitivity = 0.92, as was verified using a small set of gold standard labels).

References

Belloni, A., Chernozhukov, V., et al. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650.

Dusseldorp, E. and Van Mechelen, I. (2014). Qualitative interaction trees: a tool to identify qualitative treatment–subgroup interactions. *Statistics in medicine*, 33(2):219–237.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23.

Guo, X. and He, X. (2020). Inference on selected subgroups in clinical trials. *Journal of the American Statistical Association*, (just-accepted):1–18.

Hong, C., Liao, K. P., and Cai, T. (2019). Semi-supervised validation of multiple surrogate outcomes with application to electronic medical records phenotyping. *Biometrics*, 75(1):78–89.

Imai, K., Keele, L., and Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical science*, 25(1):51–71.

Imai, K., Ratkovic, M., et al. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470.

Kang, H., Zhang, A., Cai, T. T., and Small, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American statistical Association*, 111(513):132–144.

Mora, S., Glynn, R. J., Hsia, J., MacFadyen, J. G., Genest, J., and Ridker, P. M. (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(9):1069–1077.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9.(tlanslated and edited by dm dabrowska and tp speed, statistical science (1990), 5, 465-480). *Annals of Agricultural Sciences*, 10:1–51.

Park, M. Y. and Hastie, T. (2007). l_1 regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 69(4):659–677.

Pearl, J., Glymour, M., and Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Wang, J., He, X., and Xu, G. (2019). Debiased inference on treatment effect in a high dimensional model. *Journal of the American Statistical Association*, (just-accepted):1–000.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Annals of statistics*, 37(5A):2178.

Waters, D. D., Ho, J. E., Boekholdt, S. M., DeMicco, D. A., Kastelein, J. J., Messig, M., Breazna, A., and Pedersen, T. R. (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(2):148–152.

Windmeijer, F., Farbmacher, H., Davies, N., and Davey Smith, G. (2019). On the use of the lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association*, 114(527):1339–1350.

Würtz, P., Wang, Q., Soininen, P., Kangas, A. J., Fatemifar, G., Tynkkynen, T., Tiainen, M., Perola, M., Tillin, T., Hughes, A. D., et al. (2016). Metabolomic profiling of statin use and genetic inhibition of hmg-coa reductase. *Journal of the American college of cardiology*, 67(10):1200–1210.

Xue, H. and Pan, W. (2020). Inferring causal direction between two traits in the presence of horizontal pleiotropy with gwas summary data. *PLoS genetics*, 16(11):e1009105.

Zhang, L., Ding, X., Ma, Y., Muthu, N., Ajmal, I., Moore, J. H., Herman, D. S., and Chen, J. (2020). A maximum likelihood approach to electronic health record phenotyping using positive and unlabeled patients. *Journal of the American Medical Informatics Association*, 27(1):119–126.