This is the Pre-Published Version.

This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the American Statistical Association on 26 Feb 2019, available online: http://www.tandfonline.com/10.1080/01621459.2018.1514304.

Robust Score Tests With Missing Data in Genomics Studies

Kin Yau Wong

Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong

Donglin Zeng and D. Y. Lin

Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA

Abstract

Analysis of genomic data is often complicated by the presence of missing values, which may arise due to cost or other reasons. The prevailing approach of single imputation is generally invalid if the imputation model is misspecified. In this paper, we propose a robust score statistic based on imputed data for testing the association between a phenotype and a genomic variable with (partially) missing values. We fit a semiparametric regression model for the genomic variable against an arbitrary function of the linear predictor in the phenotype model and impute each missing value by its estimated posterior expectation. We show that the score statistic with such imputed values is asymptotically unbiased under general missing-data mechanisms, even when the imputation model is misspecified. We develop a spline-based method to estimate the semiparametric imputation model and derive the asymptotic distribution of the corresponding score statistic with a consistent variance estimator using sieve approximation theory and empirical process theory. The proposed test is computationally feasible regardless of the number of independent variables in the imputation model. We demonstrate the advantages of the proposed method over existing methods through extensive simulation studies and provide an application to a major cancer genomics study.

Keywords: Association tests; Imputation; Integrative analysis; Multiple genomics platforms; Semiparametric models; Sieve estimation.

Kin Yau Wong is Assistant Professor (E-mail: *kin-yau.wong@polyu.edu.hk*), Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong. Donglin Zeng is Professor (E-mail: *dzeng@bios.unc.edu*) and D. Y. Lin is Dennis Gillings Distinguished Professor (E-mail: *lin@bios.unc.edu*), Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA. This research was supported by the National Institutes of Health grants R01GM047845, R01HG009974, and P01CA142538.

1. INTRODUCTION

Recent technological advances have made it possible to measure multiple genomics platforms on the same set of subjects. However, constraints regarding cost and other factors prohibit measurement of all platforms on all study subjects. For example, in The Cancer Genome Atalas (TCGA) (https://cancergenome.nih.gov/), over 10,000 subjects with 33 cancer types were measured on multiple genomics platforms, including somatic mutation, copy number variation, and expressions of microRNA, mRNA, and protein, but for a substantial number of subjects, data on RNA sequencing and protein expressions were not generated. As another example, in the National Heart, Lung, and Blood Institute's Exome Sequencing Project (https://esp.gs.washington.edu/), only 7,000 subjects with specific diseases or conditions were selected for whole-exome sequencing from the tens of thousands of total subjects with genotyping array data (Lin et al. 2013). Finally, in the Trans-Omics for Precision Medicine (TOPMed) program (https://www.nhlbi.nih.gov/research/resources/nhlbi-precision-medicine-initiative/topmed) and the Genome Sequencing Program (GSP) (http:// gsp-hg.org), whole-genome sequencing data will be available on hundreds of thousands of subjects, but other genomics platforms, such as RNA sequencing, methylation, and metabolites, will be available for only a few thousand subjects through ancillary studies of specific diseases.

It is desirable to infer missing data on one genomics platform using available data from other platforms. Indeed, this has become a routine practice with genotype data, where linkage disequilibrium allows one to impute, with great accuracy, sequencing data from genotyping array data (Li et al. 2010; Auer et al. 2012). A far greater challenge is to infer missing values for a quantitative measurement, such as the expression of RNA or protein, from other quantitative measurements or from SNP genotype data due to the complex and noisy relationships among those variables (Kim et al. 2005; Torres-García et al. 2009).

Several authors have considered missing data in the context of association testing, which is of primary interest in genomics studies. Specifically, Hu et al. (2015) studied the score test based on imputed genotype data and proposed a variance estimator that properly accounts for the differential quality between observed and imputed genotypes. The method requires that the imputation is unbiased and the genotype is independent of the other variables in the phenotype model. Derkach et al. (2015) and Lawless (2016) proposed to model the variable with missing values under outcome-dependent sampling and studied the score test based on the full likelihood. Derkach et al. (2015) assumed a nonparametric model for the variable with missing values and restricted covariates to only a few possible values. Lawless (2016) assumed a full parametric missing-data model. All existing methods require unbiased imputation or correct modeling of the variable with missing values. This is difficult to achieve, especially when the number of covariates in the missing-data model is not small.

In this paper, we investigate the validity of the score test with imputed data when the missing-data mechanism may depend on the phenotype and high-dimensional covariates. In particular, we show that a condition weaker than correct specification of the missing-data model is sufficient for the score statistic to be unbiased. Based on this finding, we propose a robust score test which, unlike existing methods, preserves the type I error under general missing-data mechanisms even when the imputation model is misspecified. The proposed score statistic is based on a semiparametric model for the variable with missing values, where covariates enter the model linearly and also through a one-dimensional nonparametric function. As a result, the test is feasible with a large number of covariates in the missing-data model. The proposed methodology is applicable to all common phenotype models and encompasses continuous, binary, and right-censored phenotypes.

In Section 2, we formulate the problem, investigate the validity of the standard score test under various missing-data mechanisms, and develop the robust score test. In Section 3, we report results from simulation studies that compare the proposed and existing methods.

In Section 4, we provide an application to a dataset from TCGA. We make concluding remarks in Section 5 and relegate technical details to the Appendix.

2. METHODS

Consider a genomics study that involves phenotype Y, genomic variable of interest S, and vector of covariates $\boldsymbol{X}$. For example, Y may represent disease status, S may represent the RNA expression of a gene, and $\boldsymbol{X}$ may include genomic variables associated with S, such as the mutation status and copy number of the gene, or non-genomic variables, such as tumor stage, age, and ancestry. Let $f(\cdot; \beta S + \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X}, \boldsymbol{\zeta})$ denote the density function of Y conditional on $(S, \boldsymbol{X})$, where β and $\boldsymbol{\gamma}$ are regression parameters, and $\boldsymbol{\zeta}$ is a set of nuisance parameters that may be infinite-dimensional; this is referred to as the phenotype model. In particular, $\boldsymbol{\zeta}$ is the dispersion parameter in the generalized linear model and the baseline hazard function in the proportional hazards model. We allow S to be missing and let R indicate, by values of 1 versus 0, whether S is observed or not, respectively. Let $\boldsymbol{Z}$ be a set of predictors for S that includes $\boldsymbol{X}$, as well as variables that are not present in the phenotype model. The extra variables in $\boldsymbol{Z}$ are exogenous variables that affect Y indirectly through S and $\boldsymbol{X}$, such that $\boldsymbol{Z}$ is independent of Y conditional on $\boldsymbol{X}$ under $\beta = 0$. The observed data consist of $(Y_i, S_i R_i, R_i, \boldsymbol{Z}_i)$ for $i = 1, \ldots, n$.

We are interested in testing the null hypothesis $H_0 : \beta = 0$. Among the three common tests, namely the Wald's test, the likelihood ratio test, and the score test, the first two require fitting the model under the alternative hypothesis, which involves estimation of the conditional distribution of S given $\boldsymbol{X}$ in the presence of missing values for S. If the model for S is misspecified (which is inevitable when the dimension of $\boldsymbol{X}$ is moderately high), then the estimators of the nuisance parameters may be inconsistent, such that the resulting tests are invalid. By contrast, the score test only requires fitting the model under the null hypothesis. As a result, the score test requires fewer assumptions on the missing-data

model than the other two tests in order to yield correct type I error. Therefore, we focus on the score test in the rest of this paper.

The score statistic for β at $\beta = 0$ takes the form of $A(Y, \boldsymbol{X}; \boldsymbol{\psi})S$, where $A(Y, \boldsymbol{X}; \boldsymbol{\psi}) = \partial \log f(Y; t + \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X}, \boldsymbol{\zeta})/\partial t \mid_{t=0}$, and $\boldsymbol{\psi} = (\boldsymbol{\gamma}, \boldsymbol{\zeta})$. Note that $\mathrm{E}\{A(Y, \boldsymbol{X}; \boldsymbol{\psi}_0) \mid \boldsymbol{X}\} = 0$, where $\boldsymbol{\psi}_0 \equiv (\boldsymbol{\gamma}_0, \boldsymbol{\zeta}_0)$ is the true value of $\boldsymbol{\psi}$. This formulation includes many common models as special cases. For the linear model, $A(Y, \boldsymbol{X}; \boldsymbol{\psi}) = \sigma^{-2}(Y - \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X})$, where σ^2 is the error variance. For the logistic model, $A(Y, \boldsymbol{X}; \boldsymbol{\psi}) = Y - e^{\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X}}/(1 + e^{\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X}})$. For the proportional hazards model with right censoring, $A(Y, \boldsymbol{X}; \boldsymbol{\psi}) = \Delta - \Lambda(\widetilde{T})e^{\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X}}$, where $Y = (\widetilde{T}, \Delta)$, $\widetilde{T} = \min(T, C)$, $\Delta = I(T \leq C)$, T is the survival time of interest, C is the censoring time, $I(\cdot)$ is the indicator function, and Λ is the cumulative baseline hazard function.

We consider the score statistic based on the imputed S. We specify an imputation model of S that depends on $\boldsymbol{Z}$ and a set of parameters $\boldsymbol{\xi}$. Let $\widetilde{S}(\boldsymbol{Z}_i; \widehat{\boldsymbol{\xi}})$ be the imputed value of S_i, where $\widehat{\boldsymbol{\xi}}$ is an estimator of $\boldsymbol{\xi}$. The (normalized) imputation "score" statistic is

$$U_\beta^{\mathrm{imp}}(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\xi}}) = n^{-1/2} \sum_{i=1}^{n} A(Y_i, \boldsymbol{X}_i; \widehat{\boldsymbol{\psi}})\{R_i S_i + (1 - R_i)\widetilde{S}(\boldsymbol{Z}_i; \widehat{\boldsymbol{\xi}})\},$$

where $\widehat{\boldsymbol{\psi}} \equiv (\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\zeta}})$ is an estimator of $\boldsymbol{\psi}$ under H_0. At $\beta = 0$, the score statistic based on the full likelihood with a regression model of S on $\boldsymbol{Z}$ takes the form of U_β^{imp}. However, the proposed imputation score statistic is more general in that $\widetilde{S}$ needs not be the posterior mean of S (given the observed data) evaluated at the maximum likelihood estimator of $\boldsymbol{\xi}$. Let $\boldsymbol{\xi}^*$ be the limit of $\widehat{\boldsymbol{\xi}}$. The following proposition provides a general sufficient condition for the unbiasedness of the imputation score statistic under H_0.

Proposition 1. *Assume that there exists a projection of* $\boldsymbol{X}$*, denoted by* $\widetilde{\boldsymbol{X}}$*, such that* R *is independent of* $(S, \boldsymbol{Z})$ *conditional on* $(Y, \widetilde{\boldsymbol{X}})$ *and* $\mathrm{E}(S \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}}) = \mathrm{E}\{\widetilde{S}(\boldsymbol{Z}; \boldsymbol{\xi}^*) \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}}\}$*. Then,* $\mathrm{E}\{U_\beta^{\mathrm{imp}}(\boldsymbol{\psi}_0, \boldsymbol{\xi}^*)\} = 0$ *under* $\beta = 0$*.*

The proofs of this proposition and other technical results are provided in Appendix A.1.

Remark 1. The missing-data mechanism assumed in this proposition may arise from the extreme-tail sampling scheme, where only subjects with extreme values of Y are selected for measurements of S (Lin et al. 2013). In this case, the inverse probability weighting approach is not feasible because $P(R = 1 \mid Y)$ is zero for some subjects, whereas the imputation approach is applicable.

Remark 2. The dependence of R on $\widetilde{\boldsymbol{X}}$ may be introduced by design, where the sampling of S is performed separately at different values of $\widetilde{\boldsymbol{X}}$. In cancer genomics, $\widetilde{\boldsymbol{X}}$ may include risk factors such as tumor stage and tumor grade, and subjects with unusually high or unusually low risk may be more likely to be selected for measurements of S. In addition, $\widetilde{\boldsymbol{X}}$ may represent categories defined by (possibly continuous) demographic variables such as age. Although $\boldsymbol{X}$ may be high-dimensional and continuous, $\widetilde{\boldsymbol{X}}$ is typically a discrete, low-dimensional projection of $\boldsymbol{X}$, such that nonparametric modeling of S on $\widetilde{\boldsymbol{X}}$ is feasible.

Remark 3. The condition in Proposition 1 requires that the true and imputed S variables have the same conditional expectation given $\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}$ and $\widetilde{\boldsymbol{X}}$. This condition is trivially satisfied if S is independent of $\boldsymbol{X}$ and the imputed value has the same mean as S, as assumed by Hu et al. (2015). For the score statistic to be unbiased, we only need the expectation of the true and imputed S variables conditional on $\widetilde{\boldsymbol{X}}$ and a single index $\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}$ to be the same. This is practically achievable via nonparametric modeling of S given the low-dimensional covariates $(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}})$, even though the whole set of covariates $\boldsymbol{X}$ may be high-dimensional. If the missing-data mechanism does not depend on covariates, then $\widetilde{\boldsymbol{X}}$ is absent, such that it is only necessary to correctly model the conditional expectation of S given the single index $\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}$.

Proposition 1 implies that the imputation score statistic is unbiased under H_0 if the conditional expectation of S given a specific projection of $\boldsymbol{Z}$ (but not necessarily the full set of $\boldsymbol{Z}$) is correctly specified. To guarantee that this condition holds, we model the relationship between S and $(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}})$ nonparametrically when $\widetilde{\boldsymbol{X}}$ is discrete and takes a

small number of values. Because the regression model of S on $(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}})$ may not be very predictive, we include other components of $\boldsymbol{Z}$ in the imputation model in order to improve the imputation accuracy. Given the nonparametric function of $(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}})$, the inclusion of $\boldsymbol{Z}$ will not result in bias of the score statistic even if the imputation model is misspecified. In the sequel, we assume that the missing-data mechanism specified in Proposition 1 holds and that $\widetilde{\boldsymbol{X}}$ is discrete with possible values $(\widetilde{\boldsymbol{x}}_1, \ldots, \widetilde{\boldsymbol{x}}_L)$. For each $\widetilde{\boldsymbol{x}}_l$ $(l = 1, \ldots, L)$, we assume the working model $\mathrm{E}(S \mid \boldsymbol{Z}, \widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l) = g_l(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}) + \boldsymbol{\eta}_l^{\mathrm{T}}\widetilde{\boldsymbol{Z}}$, where g_l is unspecified, and $\widetilde{\boldsymbol{Z}}$ is a specific q-dimensional function of $\boldsymbol{Z}$ that is (asymptotically) orthogonal to $(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}})$. Let $(g_l^*, \boldsymbol{\eta}_l^*) = \arg\min_{(g_l, \boldsymbol{\eta}_l)} \mathrm{E}[R\{S - g_l(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}) - \boldsymbol{\eta}_l^{\mathrm{T}}\widetilde{\boldsymbol{Z}}\}^2 \big| \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l]$ almost surely, $\boldsymbol{\xi} = (g_1, \ldots, g_L, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_L)$, and $\boldsymbol{\xi}^* = (g_1^*, \ldots, g_L^*, \boldsymbol{\eta}_1^*, \ldots, \boldsymbol{\eta}_L^*)$. The following proposition states the unbiasedness of the resulting imputation score statistic.

Proposition 2. *If* $\widetilde{S}(\boldsymbol{Z}; \boldsymbol{\xi}^*) = \sum_{l=1}^{L} I(\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l)\{g_l^*(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}) + \boldsymbol{\eta}_l^{*\mathrm{T}}\widetilde{\boldsymbol{Z}}\}$, *then* $\mathrm{E}\{U_\beta^{\mathrm{imp}}(\boldsymbol{\psi}_0, \boldsymbol{\xi}^*)\} = 0$.

Proposition 2 motivates us to estimate g_l and $\boldsymbol{\eta}_l$ using least-squares regression with the complete observations. We propose to approximate g_l $(l = 1, \ldots, L)$ with B-spline functions of order m (De Boor 1978) and replace the true value $\boldsymbol{\gamma}_0$ by the estimator $\widehat{\boldsymbol{\gamma}}$. For simplicity of presentation, we assume the same set of fixed B-spline functions for each g_l, but we allow them to be chosen adaptively and separately for each g_l in practice. Let m and K_n be integers, such that $K_n \geq m \geq 2$, and K_n depends on the sample size n. For a set of grid points $\boldsymbol{\tau} \equiv (\tau_0, \ldots, \tau_{K_n - m + 1})$, such that $\min_{\boldsymbol{X}} \widehat{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{X} = \tau_0 < \cdots < \tau_{K_n - m + 1} = \max_{\boldsymbol{X}} \widehat{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{X}$, let $\boldsymbol{B}(\cdot) = (B_1(\cdot), \ldots, B_{K_n}(\cdot))^{\mathrm{T}}$, where B_k is the kth m-order B-spline function on $\boldsymbol{\tau}$; the grid points at the two ends have multiplicity m. For $l = 1, \ldots, L$, let

$$(\widehat{\boldsymbol{\alpha}}_l, \widehat{\boldsymbol{\eta}}_l) = \arg\min_{(\boldsymbol{\alpha}_l, \boldsymbol{\eta}_l)} \frac{1}{2} \sum_{i=1}^{n} R_i I(\widetilde{\boldsymbol{X}}_i = \widetilde{\boldsymbol{x}}_l) \Big\{ S_i - \sum_{k=1}^{K_n} \alpha_{lk} B_k(\widehat{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{X}_i) - \boldsymbol{\eta}_l^{\mathrm{T}}\widetilde{\boldsymbol{Z}}_i \Big\}^2,$$

where $\boldsymbol{\alpha}_l = (\alpha_{l1}, \ldots, \alpha_{lK_n})^{\mathrm{T}}$. Effectively, we partition the data into L strata, with each

stratum corresponding to a value of $\widetilde{\boldsymbol{x}}_l$, and we perform separate least-squares regression for each stratum using subjects with observed S. Let $\widehat{\boldsymbol{\alpha}}_l = (\widehat{\alpha}_{l1}, \ldots, \widehat{\alpha}_{lK_n})^{\mathrm{T}}$, $\widehat{g}_l = \sum_{k=1}^{K_n} \widehat{\alpha}_{lk} B_k$, and $\widehat{\boldsymbol{\xi}} = (\widehat{g}_1, \ldots, \widehat{g}_L, \widehat{\boldsymbol{\eta}}_1, \ldots, \widehat{\boldsymbol{\eta}}_L)$. The robust imputation score statistic is $U_\beta^{\mathrm{rob}}(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\xi}}; \widehat{\boldsymbol{\gamma}})$, where

$$U_\beta^{\mathrm{rob}}(\boldsymbol{\psi}, \boldsymbol{\xi}; \boldsymbol{\gamma}) = n^{-1/2} \sum_{i=1}^{n} A(Y_i, \boldsymbol{X}_i; \boldsymbol{\psi}) \Big[R_i S_i + (1 - R_i) \sum_{l=1}^{L} I(\widetilde{\boldsymbol{X}}_i = \widetilde{\boldsymbol{x}}_l) \{ g_l(\boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{X}_i) + \boldsymbol{\eta}_l^{\mathrm{T}} \widetilde{\boldsymbol{Z}}_i \} \Big],$$

and the third argument in $U_\beta^{\mathrm{rob}}(\boldsymbol{\psi}, \boldsymbol{\xi}; \boldsymbol{\gamma})$ corresponds to $\boldsymbol{\gamma}$ in the argument of g_l.

Let $\ell_\beta(\boldsymbol{\psi}, \boldsymbol{\xi}; \boldsymbol{\gamma})$ be $U_\beta^{\mathrm{rob}}(\boldsymbol{\psi}, \boldsymbol{\xi}; \boldsymbol{\gamma})$ for a single subject, $\ell_\psi(\boldsymbol{\psi})[\boldsymbol{h}_1]$ be the derivative of $\log f(Y; \boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{X}, \boldsymbol{\zeta})$ along the path $\boldsymbol{\psi} = \boldsymbol{\psi}_0 + \epsilon \boldsymbol{h}_1$, with $\boldsymbol{h}_1$ being a tangent vector for $\boldsymbol{\psi}$, $\ell_{\beta\psi}(\boldsymbol{\psi}, \boldsymbol{\xi}; \boldsymbol{\gamma})[\boldsymbol{h}_1]$ be the derivative of $\ell_\beta(\boldsymbol{\psi}, \boldsymbol{\xi}; \boldsymbol{\gamma})$ along the same path, $\ell_{\psi\psi}(\boldsymbol{\psi})[\boldsymbol{h}_1, \boldsymbol{h}_2]$ be the derivative of $\ell_\psi(\boldsymbol{\psi})[\boldsymbol{h}_1]$ along the path $\boldsymbol{\psi} = \boldsymbol{\psi}_0 + \epsilon \boldsymbol{h}_2$, with $\boldsymbol{h}_2$ being a tangent vector for $\boldsymbol{\psi}$, and $\ell_\xi(\boldsymbol{\xi})[\boldsymbol{h}_3]$ be the derivative of $R \sum_{l=1}^{L} I(\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l) \{ S - g_l(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}) - \boldsymbol{\eta}_l^{\mathrm{T}} \widetilde{\boldsymbol{Z}} \}^2 / 2$ along the path $\boldsymbol{\xi} = \boldsymbol{\xi}^* + \epsilon \boldsymbol{h}_3$, with $\boldsymbol{h}_3$ being a tangent vector for $\boldsymbol{\xi}$. Let $\mathbb{P}_n$ and P denote the empirical and true probability measures, respectively. We impose the following conditions.

(C1) For $l = 1, \ldots, L$, g_l^* and $\boldsymbol{\eta}_l^*$ are unique, and g_l^* has bounded fourth derivative.

(C2) The support of $\boldsymbol{Z}$ is bounded, and $\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}$ has a bounded continuous support. Conditional on $\boldsymbol{Z}$, S has finite second moment.

(C3) The number of knots of the B-spline functions is such that $K_n^6 n^{-1/2} \to 0$ and $K_n^7 n^{-1/2} \to \infty$ as $n \to \infty$.

(C4) At $\beta = 0$, $\|\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0\| = o_p(n^{-1/4})$ for a suitable norm, and the estimator $\widehat{\boldsymbol{\gamma}}$ satisfies

$$\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0 = \mathbb{P}_n \boldsymbol{\ell}_\gamma^*(\boldsymbol{\psi}_0) + o_p(n^{-1/2}),$$

where $\boldsymbol{\ell}_\gamma^*$ is the efficient score function of $\boldsymbol{\gamma}$, such that $P\boldsymbol{\ell}_\gamma^*(\boldsymbol{\psi}_0) = \mathbf{0}$, and $P\boldsymbol{\ell}_\gamma^*(\boldsymbol{\psi}_0)\boldsymbol{\ell}_\gamma^*(\boldsymbol{\psi}_0)^{\mathrm{T}}$

is non-zero and finite.

(C5) The functions $\ell_\beta^2(\boldsymbol{\psi}, \boldsymbol{\xi}; \boldsymbol{\gamma}_0)$, $\ell_\psi^2(\boldsymbol{\psi})[\boldsymbol{h}_1]$, $\ell_{\beta\psi}(\boldsymbol{\psi}, \boldsymbol{\xi}; \boldsymbol{\gamma}_0)[\boldsymbol{h}_1]$, and $\ell_{\psi\psi}(\boldsymbol{\psi})[\boldsymbol{h}_1, \boldsymbol{h}_2]$ are Donsker for $(\boldsymbol{\psi}, \boldsymbol{\xi})$ belonging to a neighborhood of $(\boldsymbol{\psi}_0, \boldsymbol{\xi}^*)$ and $(\boldsymbol{h}_1, \boldsymbol{h}_2)$ belonging to a bounded subset of a suitable metric space. In addition, the information operator for the phenotype model $P\ell_{\psi\psi}(\boldsymbol{\psi_0})[\cdot, \cdot]$ is invertible under the null hypothesis H_0.

Remark 4. Conditions (C1) and (C2) pertain to regularity conditions on the variable with missing values and covariates. For g_l^* and $\boldsymbol{\eta}_l^*$ to be unique, we require that $\widetilde{\boldsymbol{X}}$ and $\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}$ cannot be expressed as functions of linear terms of $\widetilde{\boldsymbol{Z}}$. In practice, we let $\widetilde{\boldsymbol{Z}}$ be a linear combination of the components of $\boldsymbol{Z}$ not present in $\widetilde{\boldsymbol{X}}$, such that $\sum_{i=1}^n \widetilde{\boldsymbol{Z}}_i \boldsymbol{Z}_i^{\mathrm{T}} \widehat{\boldsymbol{\gamma}} = \boldsymbol{0}$. Condition (C3) pertains to the rate at which the number of knots of the B-spline functions increases to infinity; particularly, the condition is satisfied with $K_n = O(n^{1/13})$. Conditions (C4) and (C5) are regularity conditions on the phenotype model, which are satisfied for common models, such as generalized linear models and proportional hazards models. For parametric models and the Cox proportional hazards model, the norm in condition (C4) is the Euclidean norm and the $\ell^\infty[0, t^*]$-norm, respectively, where t^* is the end of the study, and the metric space for $(\boldsymbol{h}_1, \boldsymbol{h}_2)$ in condition (C5) is the Euclidean space and the space of functions of bounded variation, respectively.

The asymptotic distribution of the robust imputation score statistic is given in the following theorem.

Theorem 1. *Under conditions (C1)–(C5) and $\beta = 0$, $U_\beta^{\mathrm{rob}}(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\xi}}; \widehat{\boldsymbol{\gamma}})$ is asymptotically zero-mean normal with variance*

$$V = P[\{\ell_\beta(\boldsymbol{\psi}_0, \boldsymbol{\xi}^*; \boldsymbol{\gamma}_0) - \ell_\psi(\boldsymbol{\psi}_0)[\boldsymbol{h}_\psi] - \ell_\xi(\boldsymbol{\xi}^*)[\boldsymbol{h}_\xi] - \boldsymbol{I}_\gamma(\boldsymbol{\gamma}_0, \boldsymbol{\xi}^*)^{\mathrm{T}} \boldsymbol{\ell}_\gamma^*(\boldsymbol{\psi}_0)\}^2],$$

where $\boldsymbol{h}_\psi$ solves $P\ell_{\beta\psi}(\boldsymbol{\psi}_0, \boldsymbol{\xi}^*; \boldsymbol{\gamma}_0)[\cdot] = P\ell_{\psi\psi}(\boldsymbol{\psi}_0)[\boldsymbol{h}_\psi, \cdot]$, $\boldsymbol{h}_\xi = (h_{g,1} \ldots, h_{g,L}, \boldsymbol{h}_{\eta,1}, \ldots, \boldsymbol{h}_{\eta,L})$,

such that $\boldsymbol{h}_{\eta,l} = \mathbf{0}$ *and*

$$h_{g,l}(t) = \frac{\mathrm{E}\{(1-R)I(\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l)A(Y, \boldsymbol{X}; \boldsymbol{\psi}_0) \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X} = t, \widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l\}}{\mathrm{E}\{RI(\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l) \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X} = t, \widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l\}}$$

for $l = 1, \ldots, L$,

$$\boldsymbol{I}_\gamma(\boldsymbol{\gamma}, \boldsymbol{\xi}) = \sum_{l=1}^{L} \mathrm{E}\big[I(\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l)\boldsymbol{X}\{Rg_l'(\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X})h_{g,l}(\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X}) - (1-R)A(Y, \boldsymbol{X}; \boldsymbol{\psi})g_l{}'(\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X})\}\big],$$

and f' *denotes the first derivative of* f *for any function* f.

Remark 5. The second and third terms in V are projections of the score function of $(\boldsymbol{\psi}, \boldsymbol{\xi})$, and $\boldsymbol{h}_\psi$ is the least-favorable direction of $\boldsymbol{\psi}$ for the phenotype model if the imputation model is assumed to be known. The fourth term in V is present because $\widehat{\boldsymbol{\gamma}}$, instead of the true value, is used in the imputation model. The estimator $\widehat{\boldsymbol{\gamma}}$ affects the imputation both by directly entering the imputation function $\widehat{g}_l(\widehat{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{X})$ and by involving in the estimation of $\widehat{g}_l$.

Motivated by Theorem 1, we propose an empirical variance estimator of the score statistic

$$\widehat{V} = n^{-1}\sum_{i=1}^{n}[\{\ell_{\beta,i}(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\xi}}; \widehat{\boldsymbol{\gamma}}) - \ell_{\psi,i}(\widehat{\boldsymbol{\psi}})[\widehat{\boldsymbol{h}}_\psi] - \ell_{\xi,i}(\widehat{\boldsymbol{\xi}})[\widehat{\boldsymbol{h}}_\xi] - \widehat{\boldsymbol{I}}_\gamma(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\xi}})\widehat{\boldsymbol{\ell}}_\gamma^*(\widehat{\boldsymbol{\psi}})\} - M]^2,$$

where $(\ell_{\beta,i}, \ell_{\psi,i}, \ell_{\xi,i})$ is $(\ell_\beta, \ell_\psi, \ell_\xi)$ evaluated at the observations of the ith subject, M is the sample mean of the first term in $\widehat{V}$, and $(\widehat{\boldsymbol{h}}_\psi, \widehat{\boldsymbol{h}}_\xi, \widehat{\boldsymbol{I}}_\gamma, \widehat{\boldsymbol{\ell}}_\gamma^*)$ is the empirical version of $(\boldsymbol{h}_\psi, \boldsymbol{h}_\xi, \boldsymbol{I}_\gamma, \boldsymbol{\ell}_\gamma^*)$ evaluated at $(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\xi}})$. Specifically, $\widehat{\boldsymbol{h}}_\xi$ is obtained by performing the usual linear expansion of $\widehat{\boldsymbol{\xi}}$ at $\boldsymbol{\xi}^*$, with the imputation model treated as a linear model with covariates $I(\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l)(\boldsymbol{B}(\widehat{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{X})^{\mathrm{T}}, \widetilde{\boldsymbol{Z}}^{\mathrm{T}})^{\mathrm{T}}$. The explicit form of $\widehat{\boldsymbol{h}}_\xi$ is given in the proof of Theorem 2 in Appendix A.1. We formulate the variance estimator under the linear model,

the logistic model, and the Cox proportional hazards model in Appendix A.2. The resulting score test statistic is $U_\beta^{\mathrm{rob}}(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\xi}}; \widehat{\boldsymbol{\gamma}})^2/\widehat{V}$. The validity of the robust score test is stated below.

Theorem 2. *Under conditions (C1)–(C5) and $\beta = 0$, the empirical variance estimator $\widehat{V}$ converges almost surely to the true variance V, and the test statistic $U_\beta^{\mathrm{rob}}(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\xi}}; \widehat{\boldsymbol{\gamma}})^2/\widehat{V}$ converges in distribution to the chi-square distribution with one degree of freedom.*

Remark 6. The empirical variance estimator is consistent regardless of the missing-data mechanism and the imputation model. By contrast, the standard model-based variance estimator with imputed data is generally biased if the missing-data mechanism depends on the phenotype. The bias of the standard variance estimator under generalized linear models is derived in Appendix A.3.

When the missing-data mechanism does not depend on the phenotype, the score statistic is unbiased under any imputation schemes; this result follows from the proof of Proposition 1. In this case, the proposed test is not required for bias correction, and one may wonder whether the inclusion of the B-spline terms and the stratification may lead to power loss. The comparison of power between the proposed test and the standard score test under general settings is difficult, because the power generally depends on the missing-data mechanism and high moments of S; the derivation for the power of the proposed test is given in Appendix A.4. When Y is normally distributed and S is missing completely at random, the asymptotic power of the proposed test is higher than or equal to that of the standard score test, because the imputation model with the B-spline terms, stratification, and linear predictors may have a better fit than the model with the linear predictors alone.

3. SIMULATION STUDIES

Let $\boldsymbol{X} = (X_1, X_2, X_3)^{\mathrm{T}}$, where X_1, X_2, and X_3 are independent standard normal, Bernoulli(0.5), and Binomial(2, 0.25), respectively. Let $\boldsymbol{G}$ be a vector of other covariates that are used

to predict S. In particular, $\boldsymbol{G} = (G_1, \ldots, G_4)$, where G_j $(j = 1, \ldots, 4)$ is independent Binomial$(2, 0.3)$. In cancer genomics, X_1, X_2, and X_3 may represent (standardized) age, gender, and tumor stage, respectively, and $\boldsymbol{G}$ may represent genotypes at four loci. We generated the phenotype Y using the linear predictor $r(S, \boldsymbol{X}) = \gamma_0 + \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X} + \beta S$ under the linear, logistic, and proportional hazards models. For all models, we set $\boldsymbol{\gamma} = (1, -1, 0.5)^{\mathrm{T}}$. For the linear model, we generated $Y \sim \mathrm{N}\{r(S, \boldsymbol{X}), 1\}$ with $\gamma_0 = 0$. For the logistic model, we set $\mathrm{logit}^{-1}\{P(Y = 1 \mid S, \boldsymbol{X})\} = r(S, \boldsymbol{X})$, where γ_0 was chosen such that $P(Y = 1) \approx 0.15$. For the proportional hazards model, we generated Y with the hazard function $\lambda(t \mid S, \boldsymbol{X}) = 0.5te^{r(S,\boldsymbol{X})}$ and $\gamma_0 = 0$. The censoring variable was generated independently from Unif$(0, \tau)$, where τ was chosen such that the censoring proportion was about 40%. We considered two models for S: with Model 1, $S = X_1 + X_2 + 0.3X_3 + 0.4(G_1 - G_2 + G_3 - G_4) + \mathrm{N}(0, 1)$; and with Model 2, $S = (X_1 + X_2) + 0.1(X_1 + X_2)^2 + 0.3I(X_3 = 2) + 0.4(G_1 - G_2 + G_3 - G_4) + \mathrm{N}(0, 1)$.

We considered three missing-data mechanisms. Mechanism 1 is missing completely at random, where the missing-data status is independent of other variables. For Mechanism 2, the missing-data status was generated separately for two subsets of subjects: one subset consisted of all subjects with $X_2 = 1$, and a random sample of subjects from the subset were selected for observation of S; the other subset consisted of all subjects with $X_2 = 0$, and subjects from the subset were selected for observation of S based on the phenotype. For the continuous and survival phenotypes, an equal number of subjects at the two extreme tails of the phenotype distribution were selected. For the binary phenotype, all subjects with $Y = 1$ were selected, and a fraction of subjects with $Y = 0$ were selected to attain the desired missing proportion. The missing proportion was set to be the same between the two subsets of subjects. This setting mimics a study where two datasets with different sampling schemes are combined. For Mechanism 3, four strata were defined with $\widetilde{X} \equiv \sum_{j=0}^{3} I(X_1 > z_{j/4})$ denoting the stratum number, where z_α is the α-quantile of the standard normal distribution. Subjects were selected for observation of S separately for each stratum using

the sampling scheme adopted for the second subset of subjects in Mechanism 2. The missing proportion was set to be the same across strata. This setting was designed to evaluate the sensitivity of the proposed test to misspecification of $\widetilde{\boldsymbol{X}}$.

We compared the performance of six tests: (1) the standard score test using complete data only; (2) the standard score test with missing values imputed under a linear model of S on $\boldsymbol{Z} \equiv (\boldsymbol{X}^{\mathrm{T}}, \boldsymbol{G}^{\mathrm{T}})^{\mathrm{T}}$; (3) Lawless (2016)'s score test based on the same model of S as (2); (4) Hu et al. (2015)'s score test with the imputed data of (2); (5) the proposed score test with $\widetilde{\boldsymbol{Z}}$ being that specified in Remark 4 and with stratification variable $\widetilde{X} = X_2$ for Mechanisms 1 and 2 or $\widetilde{X} = I(X_1 > 0)$ for Mechanism 3; and (6) the imputation score test with missing data imputed using a linear model of S on $\boldsymbol{Z} = (\boldsymbol{X}^{\mathrm{T}}, X_1^2, X_1X_2, I(X_3 = 2), \boldsymbol{G}^{\mathrm{T}})^{\mathrm{T}}$ and the empirical variance estimator. We refer to methods (1)–(6) as the complete-case analysis, the simple imputation method, Lawless' method, Hu's method, the proposed imputation method, and the full-model imputation method, respectively. The last method is the gold standard but is not practical because it requires correct specification of a complex missing-data model. Derkach et al. (2015)'s method was not included because it requires the covariates in the imputation model to be discrete and is identical to Lawless (2016)'s method when a linear imputation model is assumed. Note that the missing-data models used by all the methods are correct under Model 1, but only the missing-data model used by the full-model imputation method is correct under Model 2. For the proposed imputation method, we chose the degree and number of knots of the B-spline functions using five-fold cross-validation separately for each stratum. For the lth stratum, the grid point τ_k ($k = 1, \ldots, K_n - m - 2$) was set to be the empirical $k/(K_n - m + 1)$-quantile of $\widehat{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{X}_i$ among subjects with $R_i = 1$ and $\boldsymbol{X}_i = \widetilde{\boldsymbol{x}}_l$, $\tau_0 = \min_{\widetilde{\boldsymbol{X}}_i = \widetilde{x}_l} \widehat{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{X}_i$, and $\tau_{K_n - m - 1} = \max_{\widetilde{\boldsymbol{X}}_i = \widetilde{x}_l} \widehat{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{X}_i$. Lawless' and Hu's methods are not applicable to the survival phenotype.

We considered a sample size of 1,500 and missing proportions ranging from 30% to 60%. For each setting, we simulated 1,000,000 and 100,000 replicates for $\beta = 0$ and

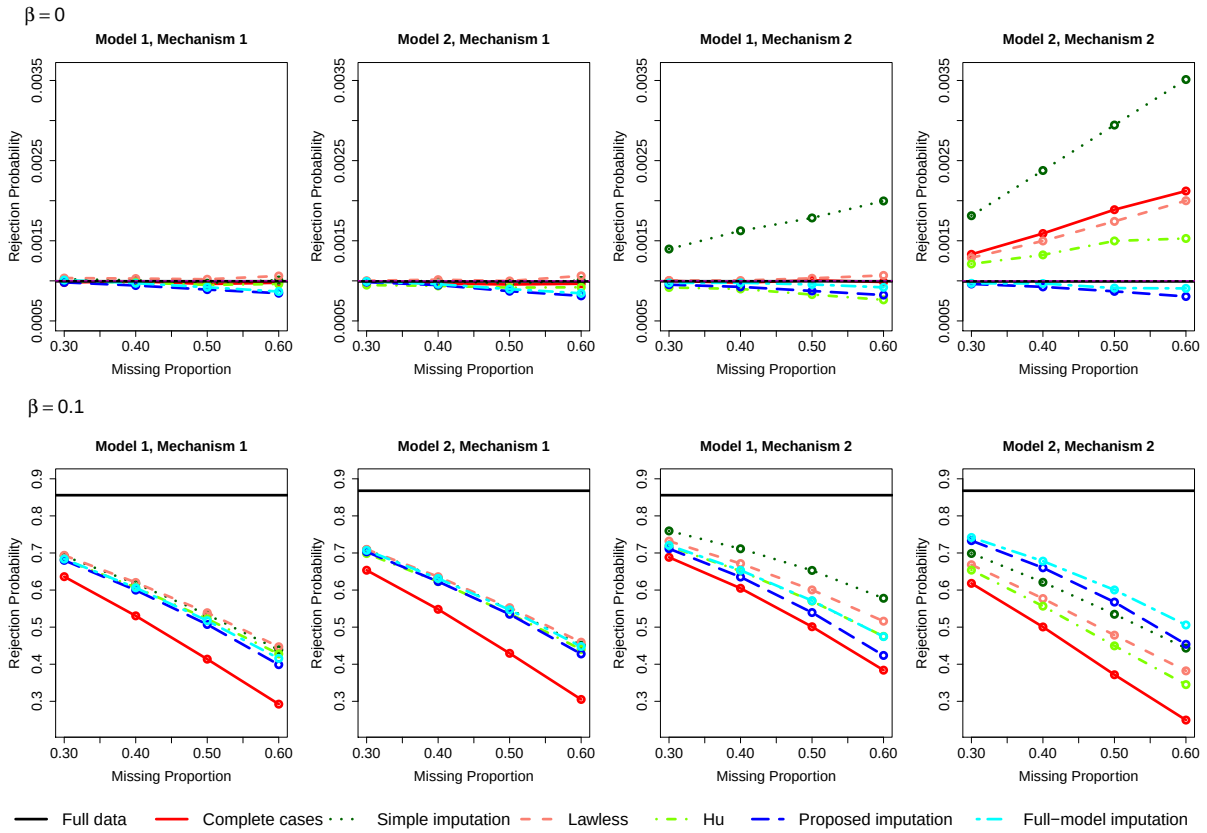

Figure 1. Rejection Probabilities for the Continuous Phenotype Under the Null and Alternative Hypotheses for Mechanisms 1 and 2.

$\beta \neq 0$, respectively. The nominal significance level was set to 10^{-3}. We plot the rejection probability against the missing proportion for the two models of S and the three missing-data mechanisms. For reference, the rejection probability of the score test based on the full data (i.e., no missing values) is also shown. For Mechanisms 1 and 2, the results of the linear model are displayed in Figure 1, and the results of the logistic and Cox proportional hazards models are displayed in Figures S1–S2 of the Supplementary Materials. For Mechanism 3, the results are shown in Figures S3–S5 of the Supplementary Materials.

Under Mechanism 1, all methods have correct type I error. Under Model 1 and Mechanism 2, the simple imputation method has inflated type I error because the variance of the

score statistic is underestimated. The complete-case analysis is also invalid except for the binary phenotype, but the type I error inflation is not as severe; the complete-case analysis for the binary phenotype has correct type I error because of the special structure of the logistic model (Prentice and Pyke 1979). Hu et al. (2015)'s variance estimator requires that both the actual and imputed S variables are independent of $\boldsymbol{X}$, which does not hold under either Model 1 or Model 2. As a result, the variance is overestimated under Mechanism 2, which leads to type I error deflation. The remaining methods have consistent variance estimators and, therefore, have correct type I error. Under Model 2 and Mechanism 2, the score statistics of the complete-case analysis and the methods based on a model of S on linear terms of $\boldsymbol{X}$ are generally biased, giving rise to type I error inflation in most cases. Hu's method exhibits type I error deflation under the logistic model because the bias of the score statistic is offset by the overestimation of the variance in this specific setting. (Because the absolute bias of the score statistic tends to infinity as $n \to \infty$, Hu's method would yield type I error inflation for large enough sample size.) The proposed imputation method is valid even though the imputation model is misspecified because the score statistic is unbiased. The full-model imputation method is also valid because the imputation model is correct. Note that the proposed imputation method and the full-model imputation method exhibit type I error deflation when the missing proportion is large. This is probably because the two methods involve a relatively large number of parameters, such that the normal approximations to the score statistics are inaccurate when the effective sample size is small.

The power of the complete-case analysis is generally low because it discards useful information. Under Model 1 or Mechanism 1, all valid methods that use the whole dataset have similar power. Under Model 2 and Mechanism 2, the full-model imputation method is the most powerful among the valid methods because a correct imputation model is assumed. However, this method cannot be used in practice because it requires knowledge

of the true relationship between S and $\boldsymbol{Z}$. The proposed imputation method is only slightly less powerful than the full-model imputation method. The bias of the score statistic of the other methods can lead to substantially low power.

Under Mechanism 3, the proposed imputation method preserves the type I error even though $\widetilde{\boldsymbol{X}}$ is misspecified. The simple imputation method underestimates the variance of the score statistic and thus yields inflated type I error, whereas Hu's method overestimates the variance and thys yields deflated type I error. For all phenotypes and models of S, the power of the proposed imputation method is similar to or higher than that of the other valid methods. Under Model 2 and the binary phenotype model, Lawless' and Hu's methods yield substantially lower power than the proposed imputation method due to the bias of the model-based score statistic. These results suggest that the proposed test is robust against misspecification of $\widetilde{\boldsymbol{X}}$, so that strata with too few data points can be collapsed.

4. REAL DATA ANALYSIS

We analyzed a dataset of patients with serous ovarian cancer from TCGA (The Cancer Genome Atlas Research Network 2011). In the study, most subjects had available genomic data, including data on DNA copy number, somatic mutation, and levels of expression of mRNA measured by microarray platforms. Only a subset of subjects had enough tissue sample left for RNA sequencing, which was introduced after the study had begun. Demographic and clinical variables, including age at diagnosis, tumor stage, tumor grade, time to tumor progression, and time to death, were available for most subjects. The median follow-up time was about 2.5 years, and roughly 30% of the patients were lost to follow-up before tumor progression or death. The data are available at http://gdac.broadinstitute.org/.

We focused on testing the association between mRNA expression, measured by RNA sequencing, and progression-free survival time. We used the fragments per kilobase of transcript per million mapped reads values for the mRNA expression variable. The number

of transcripts with RNA sequencing data was about 57,000. We considered a subset of 9,068 genes that were mutated in samples from more than five subjects. The number of subjects with available mutation, copy number, and clinical data was 407, approximately 30% of whom did not have RNA sequencing data.

We fit the Cox model for progression-free survival and included age, age squared, tumor stage, tumor grade, the interaction between age and (dichotomized) tumor stage, and the interaction between age and (dichotomized) tumor grade as covariates. Because the missing-data status is significantly associated with tumor stage (with a p-value of 4.83×10^{-6}) but not with the other covariates, we set the stratification variable $\widetilde{X}$ to be (dichotomized) tumor stage. The predictors in the imputation model of the RNA sequencing data included age, age squared, somatic mutation, copy number, and mRNA microarray expression; tumor stage and tumor grade were not included because their inclusion would render the estimation of the imputation model unstable. Microarray mRNA expression and somatic mutation were excluded from the imputation model if they were missing or too sparse. The B-spline functions were selected in the same way as in the simulation studies. For comparison, we performed the standard score test with only the complete cases and with the missing values imputed under a linear model. For further illustration, we performed the proposed test and the standard score test on the dataset with the missing proportion increased to 60%, where the RNA sequencing variables for subjects with intermediate survival or censoring time were treated as missing. The quantile-quantile plots are shown in Figure 2.

For the original dataset, the p-values from the proposed method agree with the expectation that most gene expressions are not associated with progression-free survival. The complete-case analysis and the simple imputation method yielded excessive false-positive signals because the standard variance estimates of the score statistic are smaller than the empirical variance. With extra missing data, the inflation of type I error is more severe

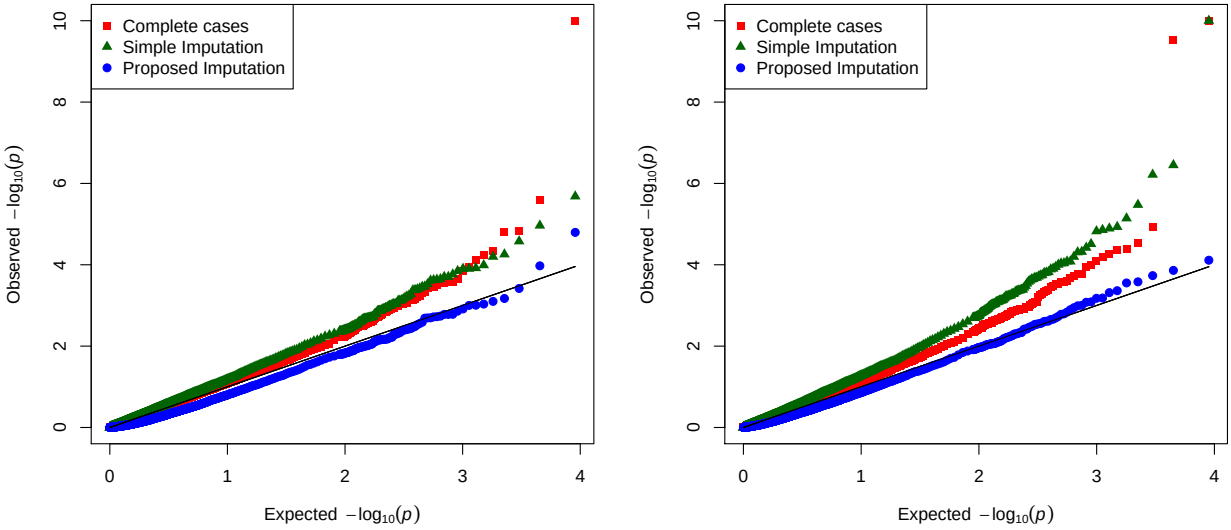

Figure 2. Quantile-Quantile Plots for the RNA-Seq Analysis of the TCGA Ovarian Cancer Data. The left plot shows the results for the original data, and the right plot shows the results with the missing proportion increased to 60%. The p-values are truncated at 10^{-10}.

for the simple imputation method, whereas the type I error is preserved by the proposed method.

The top ten genes identified by the proposed method are presented in Table 1. Several of them have been previously known to be associated with ovarian cancer, with references given in Table 1. Among these genes, the associations between progression-free survival time and the expressions of all genes except SLC4A8 are more significant under the proposed method than under the complete-case analysis. The significance levels for the associations between progression-free survival time and the expressions of WDR91, SCEL, DUSP1, VMO1, PDLIM3, DCN, CNTN4, and MARK3 are lower under the simple imputation method than under the proposed method.

Table 1. Top Genes in the RNA-Seq Analysis of the TCGA Ovarian Cancer Data.

Gene	Proposed method	Complete cases	Simple imputation	Reference
WDR91	1.60E−05	3.20E−04	2.65E−05	N/A
SLC4A8	1.06E−04	7.77E−05	1.08E−05	N/A
SCEL	3.82E−04	8.86E−04	6.29E−03	N/A
DUSP1	6.73E−04	1.26E−03	2.63E−03	Denkert et al. (2002)
VMO1	7.94E−04	9.06E−02	4.24E−03	N/A
PDLIM3	9.25E−04	1.15E−02	3.19E−03	Mougeot et al. (2006) Bignotti et al. (2007)
PLAUR	9.87E−04	5.38E−03	9.40E−04	Arend et al. (2013)
DCN	9.98E−04	1.67E−02	2.40E−03	Sherman-Baust et al. (2003)
CNTN4	1.22E−03	1.22E−01	1.32E−02	de Cristofaro et al. (2016)
MARK3	1.38E−03	2.53E−03	3.80E−04	N/A

NOTE: The top 10 genes identified by the proposed method are given in the first column, and their p-values under the proposed method, the complete-case analysis, and the simple imputation method are given in the second to fourth columns. The references for studies that have identified an association between each gene and ovarian cancer, if available, are given in the last column. "N/A" means that no prior studies that identified such an association can be found.

5. DISCUSSION

In this paper, we propose a robust score test for the association between a phenotype and a genomic variable with partially missing values. The test is based on a semiparametric model for the genomic variable, where the semiparametric component ensures that under the null hypothesis, the score statistic with imputed values is unbiased for general missing-data mechanisms and arbitrary distributions of the genomic variable. Because each nonparametric function g_l in the imputation model is a one-dimensional function of the covariates, the score test is computationally feasible with a large number of covariates provided that L is small. In addition to correcting for the bias of the score statistic, the semiparametric component results in a better fit of the imputation model, which leads to power gain even when data are missing completely at random. When the missing-data mechanism depends on the phenotype, the proposed test has correct type I error, whereas the standard score test is generally invalid. When the missing-data mechanism is independent of the phenotype, both the proposed and standard score tests have correct type I

error, but the proposed test is asymptotically more powerful.

The validity of the proposed test follows from two special properties of the score statistic under the null hypothesis. First, the phenotype model does not involve the variable with missing values, and the score statistic derived under the full likelihood coincides with the imputation score statistic. Second, the score statistic is mean zero if the expectations of the actual and imputed values conditional on a low-dimensional function of the covariates are equal. As a result, single imputation yields a valid score statistic if the expectation of the variable with missing values conditional on the low-dimensional function of the covariates is correctly specified. These two properties do not hold under the alternative hypothesis, making parameter estimation with missing data a much more challenging problem than hypothesis testing. For estimation, single imputation generally yields underestimation of standard errors (Little 1992), and a correct specification of the missing-data model is required for valid inference.

Multiple imputation (Rubin 1987) is an alternative to single imputation. To perform multiple imputation, the distribution of S conditional on the observed data, including the phenotype, has to be correctly specified. By contrast, correct specification of the distribution of S given a low-dimensional function of the covariate $\boldsymbol{X}$ is sufficient for the proposed test to be valid. Thus, multiple imputation requires much stronger assumptions than the proposed test. In fact, multiple imputation with the missing values imputed from the (correct) conditional distribution of S given $\boldsymbol{X}$ is invalid (Little 1992).

Our work can be extended in several directions. First, we have focused on a continuous genomic variable that is either exactly observed or missing. We may allow for a binary or categorical variable by incorporating the proposed semiparametric component into a generalized linear modeling framework. We may also consider genomic variables that are subject to censoring or detection limits, as in the case of metabolomics data (Yu et al. 2014). In this case, the conditional mean of the genomic variable cannot be consistently estimated

using simple least-squares estimation, and additional assumptions on the distribution of the genomic variable are necessary.

Second, it would be of interest to perform a joint test for multiple genomic variables, where each variable may have a separate pattern of missing values. This extension can be applied to many existing testing procedures that involve a multivariate score statistic, such as the sequence kernel association test for rare variants (Wu et al. 2011), tests for meta-analysis of sequencing data (Tang and Lin 2013), and the joint test for multiple genomic variables (Huang et al. 2014). Joint modeling of multiple variables with missing values is more challenging than modeling a single variable with missing value, when the pattern of the missing values for each variable do not overlap. Nevertheless, fitting a separate imputation model for each variable is not preferable, as it results in efficiency loss when the variables are correlated.

Finally, we have assumed that the missing-data mechanism depends only on the phenotype and a set of discrete covariates. The methodology can be extended to allow the missing-data mechanism to depend on continuous covariates by relating the variable with missing values to a nonparametric function of the phenotype and covariates. In addition, we may consider a missing-data mechanism that depends on a different phenotype; this scenario is common in the analysis of secondary phenotypes. In this case, the function through which the covariates affect the alternative phenotype must be estimated, and the variable with missing values should be modeled nonparametrically on that function.

APPENDIX

A.1 Proofs of Technical Results

In this section, we prove the two propositions and the two theorems put forth in the main text.

Proof of Proposition 1. The expected value of $U_\beta^{\mathrm{imp}}(\boldsymbol{\psi}_0, \boldsymbol{\xi}^*)$ is

$$
\begin{aligned}
& \mathrm{E}[A(Y, \boldsymbol{X}; \boldsymbol{\psi}_0)\{RS + (1-R)\widetilde{S}(\boldsymbol{Z}; \boldsymbol{\xi}^*)\}] \\
= \; & \mathrm{E}[A(Y, \boldsymbol{X}; \boldsymbol{\psi}_0)R\{S - \widetilde{S}(\boldsymbol{Z}; \boldsymbol{\xi}^*)\}] + \mathrm{E}\{A(Y, \boldsymbol{X}; \boldsymbol{\psi}_0)\widetilde{S}(\boldsymbol{Z}; \boldsymbol{\xi}^*)\} \\
= \; & \mathrm{E}[A(Y, \boldsymbol{X}; \boldsymbol{\psi}_0)R\mathrm{E}\{S - \widetilde{S}(\boldsymbol{Z}; \boldsymbol{\xi}^*) \mid Y, \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}}, R\}] \\
= \; & \mathrm{E}[A(Y, \boldsymbol{X}; \boldsymbol{\psi}_0)R\mathrm{E}\{S - \widetilde{S}(\boldsymbol{Z}; \boldsymbol{\xi}^*) \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}}\}] \\
= \; & \mathrm{E}[\mathrm{E}\{A(Y, \boldsymbol{X}; \boldsymbol{\psi}_0)R \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}\}\mathrm{E}\{S - \widetilde{S}(\boldsymbol{Z}; \boldsymbol{\xi}^*) \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}}\}].
\end{aligned}
$$

The second equality follows from the facts that Y is independent of $\boldsymbol{Z}$ conditional on $\boldsymbol{X}$ and that $A(Y, \boldsymbol{X}; \boldsymbol{\psi}_0)$ is a function of $(Y, \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X})$. To derive the third equality, we note that R in the conditional expectation on the third line can be omitted, because R is independent of S, $\widetilde{S}(\boldsymbol{Z}, \boldsymbol{\xi}^*)$, and $\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}$ conditional on $(Y, \widetilde{\boldsymbol{X}})$. Thus, Y in the resulting conditional expectation can be omitted, because Y is independent of S, $\widetilde{S}(\boldsymbol{Z}; \boldsymbol{\xi}^*)$, and $\widetilde{\boldsymbol{X}}$ conditional on $\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}$. We conclude that $U_\beta^{\mathrm{imp}}(\boldsymbol{\psi}_0, \boldsymbol{\xi}^*) = 0$ if $\mathrm{E}(S \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}}) = \mathrm{E}\{\widetilde{S}(\boldsymbol{Z}; \boldsymbol{\xi}^*) \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}}\}$. □

Proof of Proposition 2. First, we establish the fact that R is independent of $(S, \boldsymbol{Z})$ conditional on $(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}})$ when $\beta = 0$. To simplify the notation, for any sets of variables $\mathcal{X}$ and $\mathcal{Y}$, we use $f(\mathcal{X} \mid \mathcal{Y})$ to denote the conditional density function of $\mathcal{X}$ given $\mathcal{Y}$ evaluated at $(\mathcal{X}, \mathcal{Y})$. The conditional joint density of $(R, S, \boldsymbol{Z})$ given $(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}})$ is

$$
\begin{aligned}
& f(R, S, \boldsymbol{Z} \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}}) \\
= \; & \int f(R, S, \boldsymbol{Z}, Y \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}})\,\mathrm{d}Y \\
= \; & \int f(R \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}}, Y, S, \boldsymbol{Z}) f(Y \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}}, S, \boldsymbol{Z}) f(S, \boldsymbol{Z} \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}})\,\mathrm{d}Y \\
= \; & \int f(R \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}}, Y) f(Y \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}}) f(S, \boldsymbol{Z} \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}})\,\mathrm{d}Y \\
= \; & f(R \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}}) f(S, \boldsymbol{Z} \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}}),
\end{aligned}
$$

where the first term on the fourth line follows from the conditional independence of R to S and $\boldsymbol{Z}$ (and subsequently $\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}$) given $(Y, \widetilde{\boldsymbol{X}})$, and the second term on the same line follows from the conditional independence of Y to S and $\boldsymbol{Z}$ (and subsequently $\widetilde{\boldsymbol{X}}$) given $\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}$. The desired conditional independence follows.

Because S is independent of R conditional on $(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}})$ under $\beta = 0$,

$$\begin{aligned}
& \mathrm{E}\big[R\{S - g_l(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}) - \boldsymbol{\eta}_l^{\mathrm{T}}\widetilde{\boldsymbol{Z}}\}^2 \big| \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l\big] \\
= \; & \mathrm{E}(R \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l)\mathrm{E}\big[\{S - g_l(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}) - \boldsymbol{\eta}_l^{\mathrm{T}}\widetilde{\boldsymbol{Z}}\}^2 \big| \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l\big] \\
= \; & \mathrm{E}(R \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l)\mathrm{E}\big([S - \{g_l(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}) + \boldsymbol{\eta}_l^{\mathrm{T}}\mathrm{E}(\widetilde{\boldsymbol{Z}} \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l)\} \\
& \quad -\boldsymbol{\eta}_l^{\mathrm{T}}\{\widetilde{\boldsymbol{Z}} - \mathrm{E}(\widetilde{\boldsymbol{Z}} \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l)\}]^2 \big| \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l\big).
\end{aligned}$$

Therefore, $(g_l^*, \boldsymbol{\eta}_l^*)$ satisfies $g_l^*(t) + \boldsymbol{\eta}_l^{*\mathrm{T}}\mathrm{E}(\widetilde{\boldsymbol{Z}} \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X} = t, \widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l) = \mathrm{E}(S \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X} = t, \widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l)$. The mean of the imputed S is

$$\begin{aligned}
\mathrm{E}\{\widetilde{S}(\boldsymbol{Z}; \boldsymbol{\xi}^*) \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}}\} &= \sum_{l=1}^{L} I(\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l)\{g_l^*(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}) + \boldsymbol{\eta}_l^{*\mathrm{T}}\mathrm{E}(\widetilde{\boldsymbol{Z}} \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l)\} \\
&= \sum_{l=1}^{L} I(\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l)\mathrm{E}(S \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l) \\
&= \mathrm{E}(S \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \widetilde{\boldsymbol{X}}).
\end{aligned}$$

The desired result follows from Proposition 1. □

Before proving Theorem 1, we present the following lemma, which pertains to the consistency of $\widehat{\boldsymbol{\xi}}$.

Lemma 1. *Under conditions (C1)–(C4),* $\|\widehat{g}_l - g_l^*\|_{W^{1,\infty}} \to 0$, *and* $\|\widehat{g}_l - g_l^*\|_{L_2(X)}^2 + \|\widehat{\boldsymbol{\eta}}_l - \boldsymbol{\eta}_l^*\|_2^2 = o_p(n^{-1/2})$ *for* $l = 1, \ldots, L$, *where for any function* h *that has bounded first derivative,* $\|h\|_{W^{1,\infty}} = \|h\|_{\ell^\infty} + \|h'\|_{\ell^\infty}$, *and* $\|h\|_{L_2(X)} = [\mathrm{E}\{h(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X})^2\}]^{1/2}$.

Proof of Lemma 1. By Theorem 6.25 of Schumaker (2007), there exists $\widetilde{g}_l = \sum_{k=1}^{K_n} \widetilde{\alpha}_{lk} B_k$,

such that $\|\widetilde{g}_l - g_l^*\|_{W^{1,\infty}} \le O(K_n^{-2})$ and $\|\widetilde{g}_l - g_l^*\|_{L_2} \le O(K_n^{-7/2})$ for each $l = 1, \ldots, L$. Let $\phi_l(g_l, \boldsymbol{\eta}_l; \boldsymbol{\gamma}) = -RI(\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l)\{S - g_l(\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X}) - \boldsymbol{\eta}_l^{\mathrm{T}}\widetilde{\boldsymbol{Z}}\}^2/2$. By the definition of $\widehat{g}_l$ and $\widehat{\boldsymbol{\eta}}_l$, $\mathbb{P}_n\phi_l(\widehat{g}_l, \widehat{\boldsymbol{\eta}}_l; \widehat{\boldsymbol{\gamma}}) \ge \mathbb{P}_n\phi_l(\widetilde{g}_l, \boldsymbol{\eta}_l^*; \widehat{\boldsymbol{\gamma}})$. By the concavity of ϕ_l,

$$\mathbb{P}_n\phi_l(\epsilon\widehat{g}_l + (1-\epsilon)\widetilde{g}_l, \epsilon\widehat{\boldsymbol{\eta}}_l + (1-\epsilon)\boldsymbol{\eta}_l^*; \widehat{\boldsymbol{\gamma}}) \ge \mathbb{P}_n\phi_l(\widetilde{g}_l, \boldsymbol{\eta}_l^*; \widehat{\boldsymbol{\gamma}}),$$

where $\epsilon = (1 + \sum_{k=1}^{K_n}|\widehat{\alpha}_{lk} - \widetilde{\alpha}_{lk}| + \sum_{j=1}^{q}|\widehat{\eta}_{lj} - \eta_{lj}^*|)^{-1}$. Therefore,

$$\begin{aligned} & (\mathbb{P}_n - P)\{\phi_l(\epsilon\widehat{g}_l + (1-\epsilon)\widetilde{g}_l, \epsilon\widehat{\boldsymbol{\eta}}_l + (1-\epsilon)\boldsymbol{\eta}_l^*; \widehat{\boldsymbol{\gamma}}) - \psi(\widetilde{g}_l, \boldsymbol{\eta}_l^*; \widehat{\boldsymbol{\gamma}})\} \\ \ge\ & -P\{\phi_l(\epsilon\widehat{g}_l + (1-\epsilon)\widetilde{g}_l, \epsilon\widehat{\boldsymbol{\eta}}_l + (1-\epsilon)\boldsymbol{\eta}_l^*; \widehat{\boldsymbol{\gamma}}) - \phi_l(\widetilde{g}_l, \boldsymbol{\eta}_l^*; \widehat{\boldsymbol{\gamma}})\}. \end{aligned} \tag{A.1}$$

Note that

$$\begin{aligned} \phi_l(\widetilde{g}_l, \boldsymbol{\eta}_l^*; \widehat{\boldsymbol{\gamma}}) - \phi_l(\widetilde{g}_l, \boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0) &= R\Big\{S - \sum_{k=1}^{K_n}\widetilde{\alpha}_{lk}B_k(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{X}) - \boldsymbol{\eta}_l^{*\mathrm{T}}\widetilde{\boldsymbol{Z}}\Big\}\sum_{k=1}^{K_n}\widetilde{\alpha}_{lk}B_k{}'(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{X})(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)^{\mathrm{T}}\boldsymbol{X} \\ &\le O_p(|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0|)\max_k\|B_k{}'\|_{\ell^\infty} \\ &\le O_p(K_n n^{-1/2}), \end{aligned}$$

where the first inequality follows because $\widetilde{\alpha}_{lk}$ is bounded, and $\widetilde{\boldsymbol{\gamma}}$ is some value between $\boldsymbol{\gamma}_0$ and $\widehat{\boldsymbol{\gamma}}$. Likewise, because

$$\epsilon\widehat{\alpha}_{lk} + (1-\epsilon)\widetilde{\alpha}_{lk} = \frac{\widehat{\alpha}_{lk} - \widetilde{\alpha}_{lk}}{c_0 + |\widehat{\alpha}_{lk} - \widetilde{\alpha}_{lk}|} + \frac{1 + |\widehat{\alpha}_{lk} - \widetilde{\alpha}_{lk}|}{c_0 + |\widehat{\alpha}_{lk} - \widetilde{\alpha}_{lk}|}\widetilde{\alpha}_{lk}$$

for some $c_0 > 1$,

$$\phi_l(\epsilon\widehat{g}_l + (1-\epsilon)\widetilde{g}_l, \epsilon\widehat{\boldsymbol{\eta}}_l + (1-\epsilon)\boldsymbol{\eta}_l^*; \widehat{\boldsymbol{\gamma}}) - \phi_l(\epsilon\widehat{g}_l + (1-\epsilon)\widetilde{g}_l, \epsilon\widehat{\boldsymbol{\eta}}_l + (1-\epsilon)\boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0) \le O_p(K_n n^{-1/2}).$$

Therefore, (A.1) implies that

$$\begin{aligned} & (\mathbb{P}_n - P)\{\phi_l(\epsilon\widehat{g}_l + (1-\epsilon)\widetilde{g}_l, \epsilon\widehat{\boldsymbol{\eta}}_l + (1-\epsilon)\boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0) - \phi_l(\widetilde{g}_l, \boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0)\} + O_p(K_n n^{-1/2}) \\ \ge\ & -P\{\phi_l(\epsilon\widehat{g}_l + (1-\epsilon)\widetilde{g}_l, \epsilon\widehat{\boldsymbol{\eta}}_l + (1-\epsilon)\boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0) - \phi_l(\widetilde{g}_l, \boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0)\}. \end{aligned} \tag{A.2}$$

By Theorem 2.6.15 of van der Vaart and Wellner (1996), referred to as VW hereafter, $\{(\sum_{k=1}^{K_n} \alpha_{lk} B_k(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}) + \boldsymbol{\eta}_l^{\mathrm{T}} \widetilde{\boldsymbol{Z}})^2 : \max_{k,j}(|\alpha_{lk}|, |\eta_{lj}|) < M\}$ is a Vapnik-Chervonenkis (VC) class with VC index at most $(K_n + q)^2 + 2$ for any $M < \infty$. By Theorem 2.6.7 of VW, we show that

$$N(c_1 \|F\|_2, \mathcal{F}, L_2) \leq c_2 \{(K_n + q)^2 + 2\} (16e)^{\{(K_n+q)^2+2\}^2} \left(\frac{1}{c_1}\right)^{2\{(K_n+q)^2+1\}}$$

for any $0 \leq c_1 \leq 1$, where N is the covering number, $\mathcal{F} = \{\phi_l(g_l, \boldsymbol{\eta}_l; \boldsymbol{\gamma}_0) : g_l(\cdot) = \sum_{k=1}^{K_n} \alpha_{lk} B_k(\cdot), \max_{k,j}(|\alpha_{lk}|, |\eta_{lj}|) < M\}$, c_2 is a constant, and F is an envelope function of $\mathcal{F}$ that consists of second-order terms of $(S, \widetilde{\boldsymbol{Z}})$. Therefore, the uniform entropy of the class of functions is $O(K_n^2)$. By Theorem 2.14.1 of VW and Markov's inequality, the left-hand side of (A.2) is $O_p(K_n n^{-1/2})$.

By the linear expansion of ϕ_l at g_l^*, $P\phi_l(\widetilde{g}_l, \boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0)$ is equal to

$$\begin{aligned}
& P\phi_l(g_l^*, \boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0) - \frac{1}{2} \sum_{l=1}^{L} P(\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l) \mathrm{E}[\mathrm{E}(R \mid \boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}, \widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l) \{\widetilde{g}_l(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}) - g_l^*(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X})\}^2 \mid \widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l] \\
\leq\ & P\phi_l(g_l^*, \boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0) + O_p(\|\widetilde{g}_l - g_l^*\|_{L_2(X)}^2) \\
\leq\ & P\phi_l(g_l^*, \boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0) + O_p(K_n^{-7}),
\end{aligned}$$

where the first equality above follows because R and S are conditionally independent given $\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}$ and $\widetilde{\boldsymbol{X}}$. Likewise,

$$\begin{aligned}
& P\phi_l(\epsilon \widehat{g}_l + (1-\epsilon) \widetilde{g}_l, \epsilon \widehat{\boldsymbol{\eta}}_l + (1-\epsilon) \boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0) \\
=\ & P\phi_l(g_l^*, \boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0) - \frac{1}{2} \sum_{l=1}^{L} P(\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l) \\
& \times \mathrm{E}\{\mathrm{E}(R \mid \boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}, \widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l)[\epsilon\{\widehat{g}_l(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}) - \widetilde{g}_l(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}) + (\widehat{\boldsymbol{\eta}}_l - \boldsymbol{\eta}_l^*)^{\mathrm{T}} \widetilde{\boldsymbol{Z}}\} + \{\widetilde{g}_l(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}) - g_l^*(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X})\}]^2\}.
\end{aligned}$$

We conclude that

$$\epsilon^2\Big(\|\widehat{g}_l - \widetilde{g}_l\|^2_{L_2(X)} + \sum_{j=1}^q |\widehat{\eta}_{lj} - \eta^*_{lj}|^2\Big) \le O_p(K_n n^{-1/2}) + O_p(K_n^{-7}).$$

Because $\|\widehat{g}_l - \widetilde{g}_l\|^2_{L_2(X)} \ge c_3 \sum_{k=1}^{K_n} |\widehat{\alpha}_{lk} - \widetilde{\alpha}_{lk}|^2$ for some $c_3 > 0$,

$$\begin{aligned}
&\sum_{k=1}^{K_n} |\widehat{\alpha}_{lk} - \widetilde{\alpha}_{lk}|^2 + \sum_{j=1}^q |\widehat{\eta}_{lj} - \eta^*_{lj}|^2 \\
\le\ & \Big(1 + \sum_{k=1}^{K_n} |\widehat{\alpha}_{lk} - \widetilde{\alpha}_{lk}| + \sum_{j=1}^q |\widehat{\eta}_{lj} - \eta^*_{lj}|\Big)^2 \{O_p(K_n n^{-1/2}) + O_p(K_n^{-7})\} \\
=\ & \Big(1 + \sum_{k=1}^{K_n} |\widehat{\alpha}_{lk} - \widetilde{\alpha}_{lk}|^2 + \sum_{j=1}^q |\widehat{\eta}_{lj} - \eta^*_{lj}|^2\Big) K_n^2 \{O_p(K_n n^{-1/2}) + O_p(K_n^{-7})\},
\end{aligned}$$

which implies that $\sum_{k=1}^{K_n} |\widehat{\alpha}_{lk} - \widetilde{\alpha}_{lk}|^2 + \sum_{j=1}^q |\widehat{\eta}_{lj} - \eta^*_{lj}|^2 \le O_p(K_n^3 n^{-1/2}) + O_p(K_n^{-5})\}$. Note that

$$\begin{aligned}
\|\widehat{g}_l - \widetilde{g}_l\|_{W^{1,\infty}} &\le \max_x B_k(x) \sum_{k=1}^{K_n} |\widehat{\alpha}_{lk} - \widetilde{\alpha}_{lk}| + \max_x |B_k{}'(x)| \sum_{k=1}^{K_n} |\widehat{\alpha}_{lk} - \widetilde{\alpha}_{lk}| \\
&\le O_p\Big(K_n \sum_{k=1}^{K_n} |\widehat{\alpha}_{lk} - \widetilde{\alpha}_{lk}|\Big) = O_p(K_n^3 n^{-1/4} + K_n^{-1}).
\end{aligned}$$

Because $K_n^6 n^{-1/2} \to 0$, both $\|\widehat{g}_l - g^*_l\|_{W^{1,\infty}}$ and $\sum_{j=1}^q |\widehat{\eta}_{lj} - \eta^*_{lj}|^2$ converge to zero in probability.

To establish the rate of convergence, we replace ϵ by 1 in (A.2). The left-hand side of the resulting inequality is $o_p(n^{-1/2})$ because $\phi_l(g_l, \boldsymbol{\eta}^*_l; \boldsymbol{\gamma}_0)$ is Donsker in a neighborhood of g^*_l. By the linear expansion on the right side,

$$o_p(n^{-1/2}) \ge O_p(K_n^{-7}) + \|\widehat{g}_l - g^*_l\|^2_{L_2(X)} + \sum_{j=1}^q |\widehat{\eta}_{lj} - \eta^*_{lj}|^2.$$

The desired rate of convergence follows since $K_n^7 n^{-1/2} \to \infty$. □

Proof of Theorem 1. The imputation score statistic is

$$
\begin{aligned}
& n^{1/2}\mathbb{P}_n\ell_\beta(\widehat{\boldsymbol{\psi}},\widehat{\boldsymbol{\xi}};\widehat{\boldsymbol{\gamma}}) \\
= \; & n^{1/2}\mathbb{P}_n\ell_\beta(\boldsymbol{\psi}_0,\boldsymbol{\xi}^*;\boldsymbol{\gamma}_0)+n^{1/2}P\ell_\beta(\widehat{\boldsymbol{\psi}},\widehat{\boldsymbol{\xi}};\widehat{\boldsymbol{\gamma}})+n^{1/2}(\mathbb{P}_n-P)\{\ell_\beta(\widehat{\boldsymbol{\psi}},\widehat{\boldsymbol{\xi}};\widehat{\boldsymbol{\gamma}})-\ell_\beta(\boldsymbol{\psi}_0,\boldsymbol{\xi}^*;\boldsymbol{\gamma}_0)\} \\
= \; & n^{1/2}\mathbb{P}_n\ell_\beta(\boldsymbol{\psi}_0,\boldsymbol{\xi}^*;\boldsymbol{\gamma}_0)+n^{1/2}P\{\ell_\beta(\widehat{\boldsymbol{\psi}},\boldsymbol{\xi}^*;\widehat{\boldsymbol{\gamma}})-\ell_\beta(\boldsymbol{\psi}_0,\boldsymbol{\xi}^*;\boldsymbol{\gamma}_0)\} \\
& +n^{1/2}P\{\ell_\beta(\boldsymbol{\psi}_0,\widehat{\boldsymbol{\xi}};\boldsymbol{\gamma}_0)-\ell_\beta(\boldsymbol{\psi}_0,\boldsymbol{\xi}^*;\boldsymbol{\gamma}_0)\}+o_p(1). \qquad \text{(A.3)}
\end{aligned}
$$

The second equality above follows because the convergence rate of $(\widehat{\boldsymbol{\psi}},\widehat{\boldsymbol{\xi}})$ is at least $n^{-1/4}$ and $\ell_\beta(\boldsymbol{\psi},\boldsymbol{\xi};\boldsymbol{\gamma})$ is Donsker over a neighborhood of $(\boldsymbol{\psi}_0,\boldsymbol{\xi}^*,\boldsymbol{\gamma}_0)$. By the linear expansion, the second term of (A.3) is equal to

$$
n^{1/2}\mathrm{E}\Big\{A(Y,\boldsymbol{X};\boldsymbol{\psi}_0)(1-R)\sum_{l=1}^{L}I(\widetilde{\boldsymbol{X}}=\widetilde{\boldsymbol{x}}_l)g_0^{*\prime}(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X})\boldsymbol{X}^{\mathrm{T}}\Big\}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)+n^{1/2}P\ell_{\beta\psi}(\boldsymbol{\psi}_0,\boldsymbol{\xi}^*;\boldsymbol{\gamma}_0)[\widehat{\boldsymbol{\psi}}-\boldsymbol{\psi}_0],
$$

up to an $o_p(1)$ term. Let $\boldsymbol{I}_\gamma^{(1)}(\boldsymbol{\psi},\boldsymbol{\xi})=-\mathrm{E}\{A(Y,\boldsymbol{X};\boldsymbol{\psi})(1-R)\sum_{l=1}^{L}I(\widetilde{\boldsymbol{X}}=\widetilde{\boldsymbol{x}}_l)g_l'(\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X})\boldsymbol{X}\}$. Clearly, the first term of the above expression is $-n^{1/2}\boldsymbol{I}_\gamma^{(1)}(\boldsymbol{\psi}_0,\boldsymbol{\xi}^*)^{\mathrm{T}}\mathbb{P}_n\boldsymbol{\ell}_\gamma^*(\boldsymbol{\psi}_0)+o_p(1)$ by condition (C4). By condition (C5), $\boldsymbol{h}_\psi$ exists, and the second term of the above expression is equal to $n^{1/2}\mathbb{P}_n\ell_\psi(\boldsymbol{\psi}_0)[\boldsymbol{h}_\psi]+o_p(1)$.

The third term of (A.3) is equal to

$$
n^{1/2}\mathrm{E}\Big[A(Y,\boldsymbol{X};\boldsymbol{\psi}_0)(1-R)\sum_{l=1}^{L}I(\widetilde{\boldsymbol{X}}=\widetilde{\boldsymbol{x}}_l)\{\widehat{g}_l(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X})-g_l^*(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X})+(\widehat{\boldsymbol{\eta}}_l-\boldsymbol{\eta}_0^*)^{\mathrm{T}}\widetilde{\boldsymbol{Z}}\}\Big].
$$

Let $\ell_{g,l}(g_l,\boldsymbol{\eta}_l;\boldsymbol{\gamma})[h_{g,l}]\equiv RI(\widetilde{\boldsymbol{X}}=\widetilde{\boldsymbol{x}}_l)\{S-g_l(\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X})-\boldsymbol{\eta}_l^{\mathrm{T}}\widetilde{\boldsymbol{Z}}\}h_{g,l}(\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X})$ and $\boldsymbol{\ell}_{\eta,l}(g_l,\boldsymbol{\eta}_l;\boldsymbol{\gamma})\equiv RI(\widetilde{\boldsymbol{X}}=\widetilde{\boldsymbol{x}}_l)\{S-g_l(\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X})-\boldsymbol{\eta}_l^{\mathrm{T}}\widetilde{\boldsymbol{Z}}\}\widetilde{\boldsymbol{Z}}$ be the score function for g_l and $\boldsymbol{\eta}_l$, respectively. Let $\widetilde{h}_{g,l}$ be the B-spline approximation of $h_{g,l}$ on the same grid as $\widehat{g}_l$. By the definition of $\widehat{g}_l$ and $\widehat{\boldsymbol{\eta}}_l$,

$$
(\mathbb{P}_n-P)\{\ell_{g,l}(\widehat{g}_l,\widehat{\boldsymbol{\eta}}_l;\widehat{\boldsymbol{\gamma}})[\widetilde{h}_{g,l}]+\boldsymbol{h}_{\eta,l}^{\mathrm{T}}\boldsymbol{\ell}_{\eta,l}(\widehat{g}_l,\widehat{\boldsymbol{\eta}}_l;\widehat{\boldsymbol{\gamma}})\}=-P\{\ell_{g,l}(\widehat{g}_l,\widehat{\boldsymbol{\eta}}_l;\widehat{\boldsymbol{\gamma}})[\widetilde{h}_{g,l}]+\boldsymbol{h}_{\eta,l}^{\mathrm{T}}\boldsymbol{\ell}_{\eta,l}(\widehat{g}_l,\widehat{\boldsymbol{\eta}}_l;\widehat{\boldsymbol{\gamma}})\}.
$$

Because $\widehat{g}_l$ and $\widetilde{h}_{g,l}$ have bounded first derivatives and $\ell_{g,l}$ and $\boldsymbol{\ell}_{\eta,l}$ are differentiable, $\ell_{g,l}(g_l,\boldsymbol{\eta}_l;\boldsymbol{\gamma})[h_{g,l}]$ and $\boldsymbol{\ell}_{\eta,l}(g_l,\boldsymbol{\eta}_l;\boldsymbol{\gamma})$ are Donsker in a neighborhood of $(g_l^*,\boldsymbol{\eta}_l^*,\boldsymbol{\gamma}_0,h_{g,l})$. By

the Donsker property of the $\ell_{g,l}$ and $\boldsymbol{\ell}_{\eta,l}$ and the consistency results of Lemma 1,

$$
\begin{aligned}
& (\mathbb{P}_n - P)\{\ell_{g,l}(g_l^*, \boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0)[h_{g,l}] + \boldsymbol{h}_{\eta,l}^{\mathrm{T}} \boldsymbol{\ell}_{\eta,l}(g_l^*, \boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0)\} \\
= \; & \mathrm{E}[RI(\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l) g_l^{*\prime}(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X})\{h_{g,l}(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}) + \boldsymbol{h}_{\eta,l}^{\mathrm{T}} \widetilde{\boldsymbol{Z}}\} \boldsymbol{X}^{\mathrm{T}}](\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \\
& + \mathrm{E}[RI(\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l)\{h_{g,l}(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}) + \boldsymbol{h}_{\eta,l}^{\mathrm{T}} \widetilde{\boldsymbol{Z}}\}\{\widehat{g}_l(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}) - g_l^*(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}) + (\widehat{\boldsymbol{\eta}}_l - \boldsymbol{\eta}_l^*)^{\mathrm{T}} \widetilde{\boldsymbol{Z}}\}] + o_p(n^{-1/2}).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
& \mathrm{E}[RI(\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l)\{h_{g,l}(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}) + \boldsymbol{h}_{\eta,l}^{\mathrm{T}} \widetilde{\boldsymbol{Z}}\}\{\widehat{g}_l(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}) - g_l^*(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}) + (\widehat{\boldsymbol{\eta}}_l - \boldsymbol{\eta}_l^*)^{\mathrm{T}} \widetilde{\boldsymbol{Z}}\}] \\
= \; & (\mathbb{P}_n - P)\big\{\ell_{g,l}(g_l^*, \boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0)[h_{g,l}] + \boldsymbol{h}_{\eta,l}^{\mathrm{T}} \boldsymbol{\ell}_{\eta,l}(g_l^*, \boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0) \\
& - \mathrm{E}[RI(\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l) g_l^{*\prime}(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X})\{h_{g,l}(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}) + \boldsymbol{h}_{\eta,l}^{\mathrm{T}} \widetilde{\boldsymbol{Z}}\} \boldsymbol{X}^{\mathrm{T}}] \boldsymbol{\ell}_{\gamma}^*(\boldsymbol{\psi}_0)\big\} + o_p(n^{-1/2}).
\end{aligned}
$$

It is easy to see that $h_{g,l}$ and $\boldsymbol{h}_{\eta,l}$ solve

$$
\begin{aligned}
& \mathrm{E}[RI(\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l)\{h_{g,l}(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}) + \boldsymbol{h}_{\eta,l}^{\mathrm{T}} \widetilde{\boldsymbol{Z}}\} w(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}, \widetilde{\boldsymbol{Z}})] \\
= \; & \mathrm{E}[(1 - R) I(\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l) A(Y, \boldsymbol{X}; \boldsymbol{\psi}_0) w(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}, \widetilde{\boldsymbol{Z}})] \qquad \text{(A.4)}
\end{aligned}
$$

for all $w \in \{w(t, \widetilde{\boldsymbol{Z}}) = w_1(t) + \boldsymbol{w}_2^{\mathrm{T}} \widetilde{\boldsymbol{Z}} : w_1 \text{ has bounded fourth derivatives}, \boldsymbol{w}_2 \in \mathbb{R}^q\}$. Therefore,

$$
\begin{aligned}
& n^{1/2} P\{\ell_{\beta}(\boldsymbol{\psi}_0, \widehat{\boldsymbol{\xi}}; \boldsymbol{\gamma}_0) - \ell_{\beta}(\boldsymbol{\psi}_0, \boldsymbol{\xi}^*; \boldsymbol{\gamma}_0)\} \\
= \; & n^{1/2} \mathbb{P}_n \Big[\sum_{l=1}^{L} \{\ell_{g,l}(g_l^*, \boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0)[h_{g,l}] + \boldsymbol{h}_{\eta,l}^{\mathrm{T}} \boldsymbol{\ell}_{\eta,l}(g_l^*, \boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0)\} - \boldsymbol{I}_{\gamma}^{(2)}(\boldsymbol{\psi}_0, \boldsymbol{\xi}^*)^{\mathrm{T}} \boldsymbol{\ell}_{\gamma}^*(\boldsymbol{\psi}_0) \Big] + o_p(1) \\
= \; & n^{1/2} \mathbb{P}_n \{\ell_{\xi}(\boldsymbol{\xi}^*)[\boldsymbol{h}_{\xi}] - \boldsymbol{I}_{\gamma}^{(2)}(\boldsymbol{\psi}_0, \boldsymbol{\xi}^*)^{\mathrm{T}} \boldsymbol{\ell}_{\gamma}^*(\boldsymbol{\psi}_0)\} + o_p(1),
\end{aligned}
$$

where $\boldsymbol{I}_{\gamma}^{(2)}(\boldsymbol{\psi}, \boldsymbol{\xi}) = \sum_{l=1}^{L} \mathrm{E}[RI(\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l) g_l'(\boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{X})\{h_{g,l}(\boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{X}) + \boldsymbol{h}_{\eta,l}^{\mathrm{T}} \widetilde{\boldsymbol{Z}}\} \boldsymbol{X}]$. The desired result follows with $\boldsymbol{I}_{\gamma}(\boldsymbol{\psi}, \boldsymbol{\xi}) = \boldsymbol{I}_{\gamma}^{(1)}(\boldsymbol{\psi}, \boldsymbol{\xi}) + \boldsymbol{I}_{\gamma}^{(2)}(\boldsymbol{\psi}, \boldsymbol{\xi})$. □

Proof of Theorem 2. Under condition (C5), $\ell_{\beta}^2(\boldsymbol{\psi}, \boldsymbol{\xi})$, $\ell_{\psi}^2(\boldsymbol{\psi})[\boldsymbol{h}_1]$, $\ell_{\xi}^2(\boldsymbol{\xi})[\boldsymbol{h}_2]$, and $\boldsymbol{\ell}_{\gamma}^{*2}(\boldsymbol{\psi})$ are Glivenko-Cantelli classes over the space of $(\boldsymbol{\psi}, \boldsymbol{\xi}, \boldsymbol{h}_1, \boldsymbol{h}_2)$. Thus, the variance estimator $\widehat{V}$ is consistent if $(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\xi}}, \widehat{\boldsymbol{h}}_{\psi}, \widehat{\boldsymbol{h}}_{\xi})$ are consistent. The consistency of $(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\xi}})$ follows from

Lemma 1 and condition (C4). A consistent estimator of $\boldsymbol{h}_\psi$ can be obtained by solving $\mathbb{P}_n \ell_{\psi\psi}(\widehat{\boldsymbol{\psi}})[\boldsymbol{h}_\psi, \boldsymbol{\psi}] = \mathbb{P}_n \ell_{\beta\psi}(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\xi}}; \widehat{\boldsymbol{\gamma}})[\boldsymbol{\psi}]$ for all $\boldsymbol{\psi}$. The estimator $\widehat{\boldsymbol{h}}_\xi \equiv (\widehat{h}_{g,1}, \ldots, \widehat{h}_{g,L}, \widehat{\boldsymbol{h}}_{\eta,1}, \ldots, \widehat{\boldsymbol{h}}_{\eta,L})$ is obtained by solving the empirical version of (A.4) for the B-spline approximation of the function w. Specifically, $\widehat{h}_{g,l} = \sum_{k=1}^{K_n} \widehat{\theta}_{lk} B_k$ for $l = 1, \ldots, L$, such that $\widehat{\boldsymbol{\theta}}_l \equiv (\widehat{\theta}_{l1}, \ldots, \widehat{\theta}_{lK_n})$ and $\widehat{\boldsymbol{h}}_{\eta,l}$ solve

$$\mathbb{P}_n\big[RI(\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l) B_{lj}(\widehat{\boldsymbol{\gamma}}^{\mathrm{T}} \boldsymbol{X}) \{\widehat{h}_{g,l}(\widehat{\boldsymbol{\gamma}}^{\mathrm{T}} \boldsymbol{X}) + \boldsymbol{h}_{\eta,l}^{\mathrm{T}} \widetilde{\boldsymbol{Z}}\}\big] = \mathbb{P}_n\big\{A(Y, \boldsymbol{X}; \widehat{\boldsymbol{\psi}})(1-R) I(\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l) B_j(\widehat{\boldsymbol{\gamma}}^{\mathrm{T}} \boldsymbol{X})\big\}$$
$$\mathbb{P}_n\big[RI(\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l) \widetilde{\boldsymbol{Z}} \{\widehat{h}_{g,l}(\widehat{\boldsymbol{\gamma}}^{\mathrm{T}} \boldsymbol{X}) + \boldsymbol{h}_{\eta,l}^{\mathrm{T}} \widetilde{\boldsymbol{Z}}\}\big] = \mathbb{P}_n\big\{A(Y, \boldsymbol{X}; \widehat{\boldsymbol{\psi}})(1-R) I(\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l) \widetilde{\boldsymbol{Z}}\big\}$$

for $j = 1, \ldots, K_n$. By the Glivenko-Cantelli properties of the functions involved in the above equations and the consistency of $\widehat{\boldsymbol{\psi}}$ and the B-spline approximations, $\widehat{\boldsymbol{h}}_\xi \to \boldsymbol{h}_\xi$. The asymptotic distribution of the test statistic follows from Slutsky's theorem.

We show that the resulting projection $\ell_\xi(\boldsymbol{\xi}^*)[\widehat{\boldsymbol{h}}_\xi]$ is the same as that obtained by treating the B-spline terms as fixed covariates. Let $\boldsymbol{Z}_l^{(R)}$ and $\boldsymbol{Z}_l^{(1-R)}$ be matrices formed by stacking $(\boldsymbol{B}(\widehat{\boldsymbol{\gamma}}^{\mathrm{T}} \boldsymbol{X}_i)^{\mathrm{T}}, \widetilde{\boldsymbol{Z}}_i^{\mathrm{T}})$ together for subjects with $I(\widetilde{\boldsymbol{X}}_i = \widetilde{\boldsymbol{x}}_l, R_i = 1) = 1$ and $I(\widetilde{\boldsymbol{X}}_i = \widetilde{\boldsymbol{x}}_l, R_i = 0) = 1$, respectively, and let $\boldsymbol{A}_l^{(1-R)}$ be a vector with elements $A(Y_i, \boldsymbol{X}_i; \widehat{\boldsymbol{\psi}})$ for subjects with $\widetilde{\boldsymbol{X}}_i = \widetilde{\boldsymbol{x}}_l$ and $R_i = 0$. The estimators are

$$(\widehat{\boldsymbol{\theta}}_l^{\mathrm{T}}, \widehat{\boldsymbol{h}}_{\eta,l}^{\mathrm{T}})^{\mathrm{T}} = (\boldsymbol{Z}_l^{(R)\mathrm{T}} \boldsymbol{Z}_l^{(R)})^{-1} (\boldsymbol{Z}_l^{(1-R)\mathrm{T}} \boldsymbol{A}^{(1-R)}).$$

Let $\boldsymbol{Z}_{li}(\boldsymbol{\gamma}) = I(\widetilde{\boldsymbol{X}}_i = \widetilde{\boldsymbol{x}}_l)(\boldsymbol{B}(\boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{X}_i)^{\mathrm{T}}, \widetilde{\boldsymbol{Z}}_i^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{Z}_l$ be the generic variable. The projection of the score statistic $\ell_\xi(\boldsymbol{\xi})[\widehat{\boldsymbol{h}}_\xi]$ is equal to

$$\begin{aligned}
&\Big[S - \sum_{l=1}^{L} I(\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l)\{g_l(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}) + \boldsymbol{\eta}_l^{\mathrm{T}} \widetilde{\boldsymbol{Z}}\}\Big] \sum_{l=1}^{L} (\widehat{\boldsymbol{\theta}}_l^{\mathrm{T}}, \widehat{\boldsymbol{h}}_{\eta,l}^{\mathrm{T}}) \boldsymbol{Z}_l(\boldsymbol{\gamma}_0) \\
&= \Big[S - \sum_{l=1}^{L} I(\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l)\{g_l(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}) + \boldsymbol{\eta}_l^{\mathrm{T}} \widetilde{\boldsymbol{Z}}\}\Big] \sum_{l=1}^{L} \Big\{ \sum_{i=1}^{n} (1-R_i) I(\widetilde{\boldsymbol{X}}_i = \widetilde{\boldsymbol{x}}_l) A(Y_i, \boldsymbol{X}_i; \widehat{\boldsymbol{\psi}}) \boldsymbol{Z}_{li}(\widehat{\boldsymbol{\gamma}})^{\mathrm{T}} \Big\} \\
&\qquad \times \Big\{ \sum_{i=1}^{n} R_i I(\widetilde{\boldsymbol{X}}_i = \widetilde{\boldsymbol{x}}_l) \boldsymbol{Z}_{li}(\widehat{\boldsymbol{\gamma}}) \boldsymbol{Z}_{li}(\widehat{\boldsymbol{\gamma}})^{\mathrm{T}} \Big\}^{-1} \boldsymbol{Z}_l(\boldsymbol{\gamma}_0),
\end{aligned}$$

which is the projection of the score of $\boldsymbol{\xi}$, with $\boldsymbol{Z}_{li}$ treated as a set of fixed covariates. $\square$

A.2 Explicit Forms of Variance Estimators

In this section, we formulate the variance estimators for the linear model, the logistic model, and the Cox proportional hazards model. Let $\boldsymbol{Z}_i(\boldsymbol{\gamma}) \equiv (\boldsymbol{Z}_{1i}(\boldsymbol{\gamma})^{\mathrm{T}}, \ldots, \boldsymbol{Z}_{Li}(\boldsymbol{\gamma})^{\mathrm{T}})^{\mathrm{T}}$ denote the vector of predictors of S, with $\boldsymbol{Z}_{li}(\boldsymbol{\gamma})$ defined in the proof of Theorem 2. Let $\boldsymbol{\xi}$ denote the corresponding regression parameter and $\widehat{\boldsymbol{\xi}}$ denote the least-squares estimator of $\boldsymbol{\xi}$. Let $\boldsymbol{Z}(\boldsymbol{\gamma}) \equiv (\boldsymbol{Z}_1(\boldsymbol{\gamma})^{\mathrm{T}}, \ldots, \boldsymbol{Z}_L(\boldsymbol{\gamma})^{\mathrm{T}})^{\mathrm{T}}$ be the generic version of $\boldsymbol{Z}_i(\boldsymbol{\gamma})$. The robust imputation score statistic is

$$U_\beta^{\mathrm{rob}}(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\xi}}; \widehat{\boldsymbol{\gamma}}) = n^{-1/2} \sum_{i=1}^n \ell_{\beta,i}(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\xi}}; \widehat{\boldsymbol{\gamma}}) = n^{-1/2} \sum_{i=1}^n A(Y_i, \boldsymbol{X}_i; \widehat{\boldsymbol{\psi}}) \{R_i S_i + (1 - R_i) \widehat{\boldsymbol{\xi}}^{\mathrm{T}} \boldsymbol{Z}_i(\widehat{\boldsymbol{\gamma}})\}.$$

Let

$$\begin{aligned}
\boldsymbol{\ell}_{\xi,i}(\boldsymbol{\psi}, \boldsymbol{\xi}) &= R_i \{S_i - \boldsymbol{\xi}^{\mathrm{T}} \boldsymbol{Z}_i(\boldsymbol{\gamma})\} \boldsymbol{Z}_i(\boldsymbol{\gamma}) \\
\boldsymbol{I}_{\xi\gamma}(\boldsymbol{\psi}, \boldsymbol{\xi}) &= -\mathrm{E}\{R \boldsymbol{Z}(\boldsymbol{\gamma}) \boldsymbol{\xi}^{\mathrm{T}} \boldsymbol{Z}^{(1)}(\boldsymbol{\gamma}) \boldsymbol{X}^{\mathrm{T}}\} \\
\boldsymbol{I}_{\xi\xi}(\boldsymbol{\psi}, \boldsymbol{\xi}) &= -\mathrm{E}\{R \boldsymbol{Z}(\boldsymbol{\gamma}) \boldsymbol{Z}(\boldsymbol{\gamma})^{\mathrm{T}}\} \\
\boldsymbol{I}_{\beta\xi}(\boldsymbol{\psi}, \boldsymbol{\xi}) &= \mathrm{E}\{(1 - R) A(Y, \boldsymbol{X}; \boldsymbol{\psi}) \boldsymbol{Z}(\boldsymbol{\gamma})\},
\end{aligned}$$

where $\boldsymbol{Z}^{(1)}(\boldsymbol{\gamma}) = (\boldsymbol{Z}_1^{(1)}(\boldsymbol{\gamma})^{\mathrm{T}}, \ldots, \boldsymbol{Z}_L^{(1)}(\boldsymbol{\gamma})^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{Z}_l^{(1)}(\boldsymbol{\gamma}) = (I(\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}_l) \boldsymbol{B}_l'(\boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{X})^{\mathrm{T}}, \boldsymbol{0}_q^{\mathrm{T}})^{\mathrm{T}}$ for $l = 1, \ldots L$, the derivative $\boldsymbol{B}'$ is defined component-wise, and $\boldsymbol{0}_q$ is a q-vector of zeros.

First, we consider the linear and logistic models. For the linear model, we redefine $A(Y, \boldsymbol{X}; \boldsymbol{\psi}) = Y - \boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{X}$, where the error variance is omitted because it only acts as a scaling factor. For the two models, there is no nuisance parameter $\boldsymbol{\zeta}$, and $\boldsymbol{\psi} = \boldsymbol{\gamma}$. Let

$$\begin{aligned}
\boldsymbol{\ell}_{\gamma,i}^{\mathrm{GLM}}(\boldsymbol{\gamma}, \boldsymbol{\xi}) &= A(Y_i, \boldsymbol{X}_i; \boldsymbol{\gamma}) \boldsymbol{X}_i, \\
\boldsymbol{I}_{\gamma\gamma}^{\mathrm{GLM}}(\boldsymbol{\gamma}, \boldsymbol{\xi}) &= \mathrm{E}\{\boldsymbol{A}^{(1)}(\boldsymbol{X}; \boldsymbol{\gamma}) \boldsymbol{X}^{\mathrm{T}}\} \\
\boldsymbol{I}_{\beta\gamma}^{\mathrm{GLM}}(\boldsymbol{\gamma}, \boldsymbol{\xi}) &= \mathrm{E}[\boldsymbol{A}^{(1)}(\boldsymbol{X}; \boldsymbol{\gamma})^{\mathrm{T}} \{RS + (1 - R) \boldsymbol{\xi}^{\mathrm{T}} \boldsymbol{Z}(\boldsymbol{\gamma})\} + (1 - R) A(Y, \boldsymbol{X}; \boldsymbol{\gamma}) \boldsymbol{\xi}^{\mathrm{T}} \boldsymbol{Z}^{(1)}(\boldsymbol{\gamma}) \boldsymbol{X}^{\mathrm{T}}],
\end{aligned}$$

where $\boldsymbol{A}^{(1)}(\boldsymbol{X};\boldsymbol{\gamma}) = -\boldsymbol{X}$ for the linear model, and $\boldsymbol{A}^{(1)}(\boldsymbol{X};\boldsymbol{\gamma}) = -e^{\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X}}/(1+e^{\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X}})^2\boldsymbol{X}$ for the logistic model. In the sequel, we may omit the arguments of the above functions. The variance estimator of $U_\beta^{\mathrm{rob}}(\widehat{\boldsymbol{\psi}},\widehat{\boldsymbol{\xi}};\widehat{\boldsymbol{\gamma}})$ is

$$n^{-1}\sum_{i=1}^{n}\left\{(\ell_{\beta,i}-\overline{\ell}_\beta)-(\widehat{\boldsymbol{I}}_{\beta\gamma}^{\mathrm{GLM}}-\widehat{\boldsymbol{I}}_{\beta\xi}^{\mathrm{T}}\widehat{\boldsymbol{I}}_{\xi\xi}^{-1}\widehat{\boldsymbol{I}}_{\xi\gamma})(\widehat{\boldsymbol{I}}_{\gamma\gamma}^{\mathrm{GLM}})^{-1}(\boldsymbol{\ell}_{\gamma,i}^{\mathrm{GLM}}-\overline{\boldsymbol{\ell}}_\gamma^{\mathrm{GLM}})-\widehat{\boldsymbol{I}}_{\beta\xi}^{\mathrm{T}}\widehat{\boldsymbol{I}}_{\xi\xi}^{-1}(\boldsymbol{\ell}_{\xi,i}-\overline{\boldsymbol{\ell}}_\xi)\right\}^2\Big|_{(\boldsymbol{\gamma},\boldsymbol{\xi})=(\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\xi}})},$$

where $(\overline{\ell}_\beta,\overline{\boldsymbol{\ell}}_\psi^{\mathrm{GLM}},\overline{\boldsymbol{\ell}}_\xi)$ is the sample mean of $(\ell_{\beta,i},\boldsymbol{\ell}_{\psi,i}^{\mathrm{GLM}},\boldsymbol{\ell}_{\xi,i})$, and $(\widehat{\boldsymbol{I}}_{\gamma\gamma}^{\mathrm{GLM}},\widehat{\boldsymbol{I}}_{\xi\gamma},\widehat{\boldsymbol{I}}_{\xi\xi},\widehat{\boldsymbol{I}}_{\beta\gamma}^{\mathrm{GLM}},\widehat{\boldsymbol{I}}_{\beta\xi})$ is the empirical version of $(\boldsymbol{I}_{\gamma\gamma}^{\mathrm{GLM}},\boldsymbol{I}_{\xi\gamma},\boldsymbol{I}_{\xi\xi},\boldsymbol{I}_{\beta\gamma}^{\mathrm{GLM}},\boldsymbol{I}_{\beta\xi})$.

For the Cox proportional hazards model, let

$$\boldsymbol{\mathcal{K}}^{(r)}(t;\boldsymbol{\gamma},\boldsymbol{\xi}) = n^{-1}\sum_{j=1}^{n} I(\widetilde{T}_j\geq t)e^{\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X}_j}\begin{pmatrix}\boldsymbol{X}_j\\ R_jS_j+(1-R_j)\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{Z}_j(\boldsymbol{\gamma})\end{pmatrix}^{\otimes r}$$

for $r=0$, 1, and 2, where $\boldsymbol{a}^{\otimes 0}=1$, $\boldsymbol{a}^{\otimes 1}=\boldsymbol{a}$, and $\boldsymbol{a}^{\otimes 2}=\boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}$ for any vector $\boldsymbol{a}$. Partition $\boldsymbol{\mathcal{K}}^{(1)}=(\boldsymbol{\mathcal{K}}_1^{(1)\mathrm{T}},\mathcal{K}_2^{(1)})^{\mathrm{T}}$ with $\boldsymbol{\mathcal{K}}_1^{(1)}\in\mathbb{R}^{q\times 1}$ and $\mathcal{K}_2^{(1)}\in\mathbb{R}$ and $\boldsymbol{\mathcal{K}}^{(2)}$ into $\boldsymbol{\mathcal{K}}_{11}^{(2)}\in\mathbb{R}^{q\times q}$, $\boldsymbol{\mathcal{K}}_{12}^{(2)}\in\mathbb{R}^{q\times 1}$, $\boldsymbol{\mathcal{K}}_{21}^{(2)}\in\mathbb{R}^{1\times q}$, and $\mathcal{K}_{22}^{(2)}\in\mathbb{R}$. Let

$$\begin{aligned}
\boldsymbol{\ell}_{\gamma,i}^{\mathrm{COX}}(\boldsymbol{\psi},\boldsymbol{\xi}) &= \Delta_i\left\{\boldsymbol{X}_i-\frac{\boldsymbol{\mathcal{K}}_1^{(1)}(\widetilde{T}_i;\boldsymbol{\gamma},\boldsymbol{\xi})}{\mathcal{K}^{(0)}(\widetilde{T}_i;\boldsymbol{\gamma},\boldsymbol{\xi})}\right\}-\frac{1}{n}\sum_{j=1}^{n}\frac{I(\widetilde{T}_j\leq\widetilde{T}_i)\Delta_j e^{\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X}_j}}{\mathcal{K}^{(0)}(\widetilde{T}_j;\boldsymbol{\gamma},\boldsymbol{\xi})}\left\{\boldsymbol{X}_i-\frac{\boldsymbol{\mathcal{K}}_1^{(1)}(\widetilde{T}_j;\boldsymbol{\gamma},\boldsymbol{\xi})}{\mathcal{K}^{(0)}(\widetilde{T}_j;\boldsymbol{\gamma},\boldsymbol{\xi})}\right\}\\
\ell_{\Lambda,i}(\boldsymbol{\psi},\boldsymbol{\xi}) &= \Delta_i\frac{\mathcal{K}_2^{(1)}(\widetilde{T}_i;\boldsymbol{\gamma},\boldsymbol{\xi})}{\mathcal{K}^{(0)}(\widetilde{T}_i;\boldsymbol{\gamma},\boldsymbol{\xi})}-\Lambda(\widetilde{T}_i)e^{\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X}_i}\{R_iS_i+(1-R_i)\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{Z}_i(\boldsymbol{\gamma})\}\\
&\quad+\frac{1}{n}\sum_{j=1}^{n}\frac{I(\widetilde{T}_j\leq\widetilde{T}_i)\Delta_j e^{\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X}_j}}{\mathcal{K}^{(0)}(\widetilde{T}_j;\boldsymbol{\gamma},\boldsymbol{\xi})}\left\{R_iS_i+(1-R_i)\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{Z}_i(\boldsymbol{\gamma})-\frac{\mathcal{K}_2^{(1)}(\widetilde{T}_j;\boldsymbol{\gamma},\boldsymbol{\xi})}{\mathcal{K}^{(0)}(\widetilde{T}_j;\boldsymbol{\gamma},\boldsymbol{\xi})}\right\}\\
\boldsymbol{I}_{\beta\gamma}^{\mathrm{COX}}(\boldsymbol{\psi},\boldsymbol{\xi}) &= \mathrm{E}[\{\Delta-\Lambda(\widetilde{T})e^{\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X}}\}(1-R)\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{Z}^{(1)}(\boldsymbol{\gamma})\boldsymbol{X}^{\mathrm{T}}]+\boldsymbol{v}_{12}(\boldsymbol{\gamma},\boldsymbol{\xi})^{\mathrm{T}}\\
\boldsymbol{I}_{\gamma\gamma}^{\mathrm{COX}}(\boldsymbol{\psi},\boldsymbol{\xi}) &= \boldsymbol{v}_{11}(\boldsymbol{\gamma},\boldsymbol{\xi}),
\end{aligned}$$

where

$$\boldsymbol{v}_{jk}(\boldsymbol{\gamma};\boldsymbol{\xi}) = -\mathrm{E}\left[\Delta\left\{\frac{\boldsymbol{\mathcal{K}}_{jk}^{(2)}(\widetilde{T};\boldsymbol{\gamma},\boldsymbol{\xi})^{\mathrm{T}}}{\mathcal{K}^{(0)}(\widetilde{T};\boldsymbol{\gamma},\boldsymbol{\xi})}-\frac{\boldsymbol{\mathcal{K}}_j^{(1)}(\widetilde{T};\boldsymbol{\gamma},\boldsymbol{\xi})\boldsymbol{\mathcal{K}}_k^{(1)}(\widetilde{T};\boldsymbol{\gamma},\boldsymbol{\xi})^{\mathrm{T}}}{\mathcal{K}^{(0)}(\widetilde{T};\boldsymbol{\gamma},\boldsymbol{\xi})^2}\right\}\right]\quad\text{for } j,\ k=1 \text{ or } 2.$$

The variance estimator of $U_\beta^{\rm rob}(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\xi}}; \widehat{\boldsymbol{\gamma}})$ is

$$n^{-1}\sum_{i=1}^{n}\left\{(\ell_{\beta,i}-\overline{\ell}_\beta)-(\widehat{\boldsymbol{I}}_{\beta\gamma}^{\rm COX}-\widehat{\boldsymbol{I}}_{\beta\xi}\widehat{\boldsymbol{I}}_{\xi\xi}^{-1}\widehat{\boldsymbol{I}}_{\xi\gamma})(\widehat{\boldsymbol{I}}_{\gamma\gamma}^{\rm COX})^{-1}(\boldsymbol{\ell}_{\gamma,i}^{\rm COX}-\overline{\boldsymbol{\ell}}_\gamma^{\rm COX})-(\ell_{\Lambda,i}-\overline{\ell}_\Lambda)-\widehat{\boldsymbol{I}}_{\beta\xi}\widehat{\boldsymbol{I}}_{\xi\xi}^{-1}(\boldsymbol{\ell}_{\xi,i}-\overline{\boldsymbol{\ell}}_\xi)\right\}^2,$$

with $(\boldsymbol{\psi}, \boldsymbol{\xi})$ evaluated at $(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\xi}})$, where $(\overline{\boldsymbol{\ell}}_\gamma^{\rm COX}, \overline{\ell}_\Lambda^{\rm COX})$ is the sample mean of $(\boldsymbol{\ell}_{\gamma,i}^{\rm COX}, \ell_{\Lambda,i}^{\rm COX})$, and $(\widehat{\boldsymbol{I}}_{\beta\gamma}^{\rm COX}\widehat{\boldsymbol{I}}_{\gamma\gamma}^{\rm COX})$ is the empirical version of $(\boldsymbol{I}_{\beta\gamma}^{\rm COX}\boldsymbol{I}_{\gamma\gamma}^{\rm COX})$.

A.3 Bias of the Standard Variance Estimator

In this section, we evaluate the standard model-based variance estimator based on imputed data. Consider a generalized linear model with no nuisance parameter. The information matrix of $(\beta, \boldsymbol{\gamma})$ under $\beta = 0$ is

$$\begin{pmatrix} B(\boldsymbol{X};\boldsymbol{\gamma})S^2 & B(\boldsymbol{X};\boldsymbol{\gamma})S\boldsymbol{X}^{\rm T} \\ B(\boldsymbol{X};\boldsymbol{\gamma})S\boldsymbol{X} & B(\boldsymbol{X};\boldsymbol{\gamma})\boldsymbol{X}\boldsymbol{X}^{\rm T} \end{pmatrix},$$

where $B(\boldsymbol{X};\boldsymbol{\gamma}) = {\rm Var}\{A(Y, \boldsymbol{X};\boldsymbol{\gamma}) \mid \boldsymbol{X}\}$. The limit of the model-based variance estimator with the imputed data is equal to

$$\begin{aligned}\widehat{V}^{\rm std} &= {\rm E}[B(\boldsymbol{X};\boldsymbol{\gamma}_0)\{\widetilde{S}^2 - {\rm E}(\widetilde{S}\boldsymbol{X})^{\rm T}{\rm E}(\boldsymbol{X}\boldsymbol{X}^{\rm T})^{-1}\widetilde{S}\boldsymbol{X}\}] \\ &= {\rm E}\{B(\boldsymbol{X};\boldsymbol{\gamma}_0)\widetilde{S}^2\} - {\rm E}\{B(\boldsymbol{X};\boldsymbol{\gamma}_0)S\boldsymbol{X}^{\rm T}\}{\rm E}\{B(\boldsymbol{X};\boldsymbol{\gamma}_0)\boldsymbol{X}\boldsymbol{X}^{\rm T}\}^{-1}{\rm E}\{B(\boldsymbol{X};\boldsymbol{\gamma}_0)S\boldsymbol{X}\},\end{aligned}$$

where $\widetilde{S}$ is equal to S if $R = 1$ and is equal to $\widetilde{S}(\boldsymbol{Z};\boldsymbol{\xi}^*)$ otherwise.

We derive the bias of the model-based variance estimator under a correct imputation model, i.e., ${\rm E}(S \mid \boldsymbol{X}) = {\rm E}\{\widetilde{S}(\boldsymbol{Z};\boldsymbol{\xi}^*) \mid \boldsymbol{X}\}$, and a balanced sampling scheme of S, i.e., ${\rm E}\{A(Y, \boldsymbol{X};\boldsymbol{\gamma}_0)R \mid \boldsymbol{X}\} = 0$. Assume that $\sup_{\boldsymbol{Z}}|\widetilde{S}(\boldsymbol{Z};\widehat{\boldsymbol{\xi}}) - \widetilde{S}(\boldsymbol{Z};\boldsymbol{\xi}^*)| = O_p(n^{-1/2})$. In this case, the imputation score statistic is

$$\begin{aligned} & U_\beta^{\rm imp}(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\xi}}) \\ = \; & n^{-1/2}\sum_{i=1}^{n} A(Y_i, \boldsymbol{X}_i;\widehat{\boldsymbol{\gamma}})\{R_iS_i + (1-R_i)\widetilde{S}(\boldsymbol{Z}_i;\widehat{\boldsymbol{\xi}})\}\end{aligned}$$

$$
\begin{aligned}
&= n^{-1/2}\sum_{i=1}^{n} A(Y_i, \boldsymbol{X}_i; \boldsymbol{\gamma}_0)\widetilde{S}_i - B(\boldsymbol{X}_i; \boldsymbol{\gamma}_0)\widetilde{S}_i \boldsymbol{X}_i^{\mathrm{T}}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \\
&\quad + A(Y_i, \boldsymbol{X}_i; \boldsymbol{\gamma}_0)(1 - R_i)\{\widetilde{S}(\boldsymbol{Z}; \widehat{\boldsymbol{\xi}}) - \widetilde{S}(\boldsymbol{Z}; \boldsymbol{\xi}^*)\} + o_p(1) \\
&= n^{-1/2}\sum_{i=1}^{n} A(Y_i, \boldsymbol{X}_i; \boldsymbol{\gamma}_0)[\widetilde{S}_i - \mathrm{E}\{B(\boldsymbol{X}; \boldsymbol{\gamma}_0)\widetilde{S}\boldsymbol{X}^{\mathrm{T}}\}\mathrm{E}\{B(\boldsymbol{X}; \boldsymbol{\gamma}_0)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\}^{-1}\boldsymbol{X}_i] + o_p(1),
\end{aligned}
$$

where the third equality follows from the rate of convergence of $\widetilde{S}(\boldsymbol{Z}; \widehat{\boldsymbol{\xi}})$ and the balanced sampling scheme. Therefore, the asymptotic variance of the imputation score statistic is

$$
\mathrm{E}\{A(Y, \boldsymbol{X}; \boldsymbol{\gamma}_0)^2\widetilde{S}^2\} - \mathrm{E}\{B(\boldsymbol{X}; \boldsymbol{\gamma}_0)S\boldsymbol{X}^{\mathrm{T}}\}\mathrm{E}\{B(\boldsymbol{X}; \boldsymbol{\gamma}_0)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\}^{-1}\mathrm{E}\{B(\boldsymbol{X}; \boldsymbol{\gamma}_0)S\boldsymbol{X}\}.
$$

The bias of the model-based estimator is

$$
\mathrm{E}\{B(\boldsymbol{X}; \boldsymbol{\gamma}_0)\widetilde{S}^2\} - \mathrm{E}\{A(Y, \boldsymbol{X}; \boldsymbol{\gamma}_0)^2\widetilde{S}^2\}. \tag{A.5}
$$

Let $v_{Y,j}(\boldsymbol{X}) = \mathrm{Var}\{A(Y, \boldsymbol{X}; \boldsymbol{\gamma}) \mid R = j, \boldsymbol{X}\}$, $p_R(\boldsymbol{X}) = P(R = 1 \mid \boldsymbol{X})$, $v_S(\boldsymbol{X}) = \mathrm{E}(S^2 \mid \boldsymbol{X})$, and $v_{\widetilde{S}}(\boldsymbol{X}) = \mathrm{E}\{\widetilde{S}(\boldsymbol{Z}; \boldsymbol{\xi}^*)^2 \mid \boldsymbol{X}\}$. The first term of (A.5) is $\mathrm{E}\{B(\boldsymbol{X}; \boldsymbol{\gamma}_0)p_R(\boldsymbol{X})v_S(\boldsymbol{X})\} + \mathrm{E}[B(\boldsymbol{X}; \boldsymbol{\gamma}_0)\{1 - p_R(\boldsymbol{X})\}v_{\widetilde{S}}(\boldsymbol{X})]$. By the definition of $B(\boldsymbol{X}; \boldsymbol{\gamma}_0)$,

$$
\begin{aligned}
B(\boldsymbol{X}; \boldsymbol{\gamma}_0) &= \mathrm{Var}\{A(Y, \boldsymbol{X}; \boldsymbol{\gamma}_0) \mid \boldsymbol{X}\} \\
&= \mathrm{E}[\mathrm{Var}\{A(Y, \boldsymbol{X}; \boldsymbol{\gamma}_0) \mid \boldsymbol{X}, R\} \mid \boldsymbol{X}] + \mathrm{Var}[\mathrm{E}\{A(Y, \boldsymbol{X}; \boldsymbol{\gamma}_0) \mid \boldsymbol{X}, R\} \mid \boldsymbol{X}] \\
&= p_R(\boldsymbol{X})v_{Y,1}(\boldsymbol{X}) + \{1 - p_R(\boldsymbol{X})\}v_{Y,0}(\boldsymbol{X}).
\end{aligned}
$$

The second term of (A.5) is

$$
\begin{aligned}
&\mathrm{E}[A(Y, \boldsymbol{X}; \boldsymbol{\gamma}_0)^2\{RS + (1 - R)\widetilde{S}(\boldsymbol{Z}; \boldsymbol{\xi}^*)\}^2] \\
&= \mathrm{E}\{A(Y, \boldsymbol{X}; \boldsymbol{\gamma}_0)^2 RS^2\} + \mathrm{E}\{A(Y, \boldsymbol{X}; \boldsymbol{\gamma}_0)^2(1 - R)\widetilde{S}(\boldsymbol{Z}; \boldsymbol{\xi}^*)^2\} \\
&= \mathrm{E}[v_{Y,1}(\boldsymbol{X})p_R(\boldsymbol{X})v_S(\boldsymbol{X})\} + \mathrm{E}\{v_{Y,0}(\boldsymbol{X})\{1 - p_R(\boldsymbol{X})\}v_{\widetilde{S}}(\boldsymbol{X})].
\end{aligned}
$$

To see that (A.5) is non-zero in general, consider the simple case of $X = 1$, $v_{Y,1} > v_{Y,0}$, and $v_S > v_{\widetilde{S}}$, such that there is no covariate, the variance of the phenotype is larger among

subjects with observed S, and the variance of the true S is larger than that of the imputed S. By Chebyshev's sum inequality, the true variance $\mathrm{E}\{A(Y,\boldsymbol{X};\boldsymbol{\gamma}_0)^2\widetilde{S}^2\}$ is strictly larger than the limit of the model-based variance estimator $\mathrm{E}\{B(\boldsymbol{X};\boldsymbol{\gamma}_0)^2\widetilde{S}^2\}$. By contrast, if the missing-data mechanism does not depend on Y, then $v_{Y,1}$ and $v_{Y,0}$ are equal. As a result, (A.5) is equal to zero, and the model-based variance estimator is consistent.

A.4 Evaluation of Power

In this section, we evaluate the power of the imputation score test under the linear model: $Y = \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X} + \beta S + \mathrm{N}(0,\sigma^2)$. Assume the same imputation model for S as in Section A.2 with a general predictor $\boldsymbol{Z}(\boldsymbol{\gamma})$. Assume also that $\boldsymbol{X}$ is contained in $\boldsymbol{Z}(\boldsymbol{\gamma})$ and that $\mathrm{E}(S \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X},\widetilde{\boldsymbol{X}}) = \mathrm{E}\{\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0) \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X},\widetilde{\boldsymbol{X}}\}$, such that the score statistic is unbiased. The missing-data status can be expressed as $R = I\{(Y,\widetilde{\boldsymbol{X}}) \in \Omega(\omega)\}$, where for every fixed ω, $\Omega(\omega)$ is a deterministic subset of the space of $(Y,\widetilde{\boldsymbol{X}})$, and ω is a random variable that is independent of $(Y,\boldsymbol{Z}(\boldsymbol{\gamma}),S)$. Adopting the notation introduced in Section A.2, the score statistic can be expanded as $n^{-1/2}\sum_{i=1}^n \boldsymbol{\Psi}(\ell_{\beta,i},\boldsymbol{\ell}_{\gamma,i}^{\mathrm{T}},\boldsymbol{\ell}_{\xi,i}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\Psi} = (1,(\boldsymbol{I}_{\beta\gamma} - \boldsymbol{I}_{\beta\xi}^{\mathrm{T}}\boldsymbol{I}_{\xi\xi}^{-1}\boldsymbol{I}_{\xi\gamma})\boldsymbol{I}_{\gamma\gamma}^{-1}, -\boldsymbol{I}_{\beta\xi}\boldsymbol{I}_{\xi\xi}^{-1})$, $\boldsymbol{\ell}_{\gamma,i} = \boldsymbol{\ell}_{\gamma,i}^{\mathrm{GLM}}$, $\boldsymbol{I}_{\gamma\gamma} = \boldsymbol{I}_{\gamma\gamma}^{\mathrm{GLM}}$, and $\boldsymbol{I}_{\beta\gamma} = \boldsymbol{I}_{\beta\gamma}^{\mathrm{GLM}}$.

Under the contiguous alternative of $\beta_n = n^{-1/2}b$ for some fixed b,

$$\begin{aligned} n^{-1/2}\mathrm{E}\Big\{\sum_{i=1}^n \ell_{\beta,i}(\boldsymbol{\gamma}_0,\boldsymbol{\xi}^*;\boldsymbol{\gamma}_0)\Big\} &= n^{1/2}\mathrm{E}[(Y-\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X})\{RS+(1-R)\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}] \\ &= n^{1/2}\mathrm{E}[\varepsilon R\{S-\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}] + b\mathrm{E}[S\{RS+(1-R)\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}], \end{aligned}$$

where $\varepsilon = Y - \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X} - \beta_n S$. The first expectation on the far right side of the above expression is

$$\begin{aligned} & \mathrm{E}[\varepsilon I\{(Y,\widetilde{\boldsymbol{X}}) \in \Omega(\omega)\}\{S-\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}] \\ = \; & \mathrm{E}[\varepsilon\{S-\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}P\{(Y,\widetilde{\boldsymbol{X}}) \in \Omega(\omega) \mid \varepsilon, S, \boldsymbol{Z}(\boldsymbol{\gamma}_0)\}] \\ = \; & \mathrm{E}[\varepsilon\{S-\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}P\{(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}+\varepsilon,\widetilde{\boldsymbol{X}}) \in \Omega(\omega) \mid \varepsilon, \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X},\widetilde{\boldsymbol{X}}\}] \end{aligned}$$

$$+\beta_n \mathrm{E}\Big[\varepsilon\{S-\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}S\frac{\partial}{\partial t}P\{(t+\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}+\varepsilon,\widetilde{\boldsymbol{X}})\in\Omega(\omega)\mid\varepsilon,\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X},\widetilde{\boldsymbol{X}}\}|_{t=0}\Big]+o(1).$$

The first term of the right side of the last equality is zero by assumption. Thus,

$$\begin{aligned}
& n^{-1/2}\mathrm{E}\Big\{\sum_{i=1}^{n}\ell_{\beta,i}(\boldsymbol{\gamma}_0,\boldsymbol{\xi}^*;\boldsymbol{\gamma}_0)\Big\} \\
\to\ & b\mathrm{E}\Big[\varepsilon\{S-\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}S\frac{\partial}{\partial t}P\{(t+\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}+\varepsilon,\widetilde{\boldsymbol{X}})\in\Omega(\omega)\mid\varepsilon,\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X},\widetilde{\boldsymbol{X}}\}|_{t=0}\Big] \\
& \quad +b\mathrm{E}[S\{RS+(1-R)\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}] \\
\equiv\ & b\mu_\beta.
\end{aligned}$$

Simple algebraic manipulation yields $n^{-1/2}\mathrm{E}\{\sum_{i=1}^{n}\boldsymbol{\ell}_{\gamma,i}(\boldsymbol{\gamma}_0,\boldsymbol{\xi}^*)\} = b\mathrm{E}(S\boldsymbol{X}) \equiv b\boldsymbol{\mu}_\gamma$. In addition,

$$\begin{aligned}
& n^{-1/2}\mathrm{E}\Big\{\sum_{i=1}^{n}\boldsymbol{\ell}_{\xi,i}(\boldsymbol{\gamma}_0,\boldsymbol{\xi}^*)\Big\} \\
=\ & n^{1/2}\mathrm{E}[I\{(Y,\widetilde{\boldsymbol{X}})\in\Omega(\omega)\}\{S-\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}\boldsymbol{Z}(\boldsymbol{\gamma}_0)] \\
=\ & n^{1/2}\mathrm{E}[P\{(Y,\widetilde{\boldsymbol{X}})\in\Omega(\omega)\mid S,\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}\{S-\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}\boldsymbol{Z}(\boldsymbol{\gamma}_0)] \\
\to\ & b\mathrm{E}\Big[\frac{\partial}{\partial t}P\{(t+\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}+\varepsilon,\widetilde{\boldsymbol{X}})\in\Omega(\omega)\mid\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X},\widetilde{\boldsymbol{X}}\}|_{t=0}S\{S-\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\Big]\equiv b\boldsymbol{\mu}_\xi.
\end{aligned}$$

Therefore, the asymptotic distribution of the score test statistic is non-central chi-square with non-centrality parameter

$$\mathcal{C}=\frac{b^2\{\boldsymbol{\Psi}(\mu_\beta,\boldsymbol{\mu}_\gamma^{\mathrm{T}},\boldsymbol{\mu}_\xi^{\mathrm{T}})^{\mathrm{T}}\}^2}{\boldsymbol{\Psi}\mathrm{E}\{(\ell_{\beta,i},\boldsymbol{\ell}_{\gamma,i}^{\mathrm{T}},\boldsymbol{\ell}_{\xi,i}^{\mathrm{T}})^{\otimes 2}\}\boldsymbol{\Psi}^{\mathrm{T}}}.$$

Because the conditional distribution of S given $\boldsymbol{Z}(\boldsymbol{\gamma}_0)$ may be misspecified, $\{S-\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}$ in $\boldsymbol{\ell}_{\xi,i}(\boldsymbol{\gamma}_0,\boldsymbol{\xi}^*)$ is generally dependent of $\boldsymbol{Z}(\boldsymbol{\gamma}_0)$. Thus, the non-centrality parameter is a function of high moments of $\boldsymbol{Z}(\boldsymbol{\gamma}_0)$, and it is difficult to evaluate the power

for different choices of $\boldsymbol{Z}(\boldsymbol{\gamma}_0)$. Consider the case that S is missing completely at random. In this case, $\boldsymbol{I}_{\beta\gamma} = -\mathrm{E}[\{RS + (1-R)\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}\boldsymbol{X}^{\mathrm{T}}]$, and $\boldsymbol{I}_{\beta\xi} = \boldsymbol{0}$. The denominator of $\mathcal{C}$ is

$$
\begin{aligned}
& \mathrm{Var}(\ell_{\beta,i} - \boldsymbol{I}_{\beta\gamma}\boldsymbol{I}_{\gamma\gamma}^{-1}\boldsymbol{\ell}_{\gamma,i}) \\
= \; & \mathrm{E}[\varepsilon^2\{RS + (1-R)\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0) - \boldsymbol{I}_{\beta\gamma}\boldsymbol{I}_{\gamma\gamma}^{-1}\boldsymbol{X}\}^2] \\
= \; & \sigma^2\mathrm{E}\{RS + (1-R)\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0) - \mathrm{E}[\{RS + (1-R)(\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}\boldsymbol{X}]\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\boldsymbol{X}\}^2 \\
= \; & \sigma^2\mathrm{E}[\mathcal{P}_{\boldsymbol{X}}^{\perp}\{RS + (1-R)\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}]^2,
\end{aligned}
$$

where $\mathcal{P}_{\boldsymbol{X}}^{\perp}$ denotes the projection onto the orthogonal space of $\boldsymbol{X}$, i.e., $\mathcal{P}_{\boldsymbol{X}}^{\perp}T = T - \boldsymbol{X}^{\mathrm{T}}\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\mathrm{E}(\boldsymbol{X}T)$ for any random variable T. The numerator of $\mathcal{C}$ is

$$
\begin{aligned}
& b^2(\mu_\beta - \boldsymbol{I}_{\beta\gamma}\boldsymbol{I}_{\gamma\gamma}^{-1}\boldsymbol{\mu}_\gamma)^2 \\
= \; & b^2(\mathrm{E}\{RS^2 + (1-R)S\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\} - \mathrm{E}(S\boldsymbol{X}^{\mathrm{T}})\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\mathrm{E}[\{RS + (1-R)\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}\boldsymbol{X}])^2 \\
= \; & b^2(\mathrm{E}[R\{S^2 - \mathrm{E}(S\boldsymbol{X}^{\mathrm{T}})\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\boldsymbol{X}S\} \\
& \quad +(1-R)\{S\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0) - \mathrm{E}(S\boldsymbol{X}^{\mathrm{T}})\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\boldsymbol{X}\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}])^2 \\
= \; & b^2[p_R\mathrm{E}(\mathcal{P}_{\boldsymbol{X}}^{\perp}S)^2 + (1-p_R)\mathrm{Cov}\{\mathcal{P}_{\boldsymbol{X}}^{\perp}S, \mathcal{P}_{\boldsymbol{X}}^{\perp}\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}]^2 \\
= \; & b^2[p_R\mathrm{E}(\mathcal{P}_{\boldsymbol{X}}^{\perp}S)^2 + (1-p_R)\mathrm{E}\{\mathcal{P}_{\boldsymbol{X}}^{\perp}\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}^2]^2,
\end{aligned}
$$

where $p_R = P(R=1)$, and the last equality follows from the definition of $\boldsymbol{\xi}^*$ and the fact that $\boldsymbol{X}$ is contained in $\boldsymbol{Z}(\boldsymbol{\gamma}_0)$. As a result,

$$
\mathcal{C} = \frac{b^2}{\sigma^2} \times \frac{[p_R\mathrm{E}(\mathcal{P}_{\boldsymbol{X}}^{\perp}S)^2 + (1-p_R)\mathrm{E}\{\mathcal{P}_{\boldsymbol{X}}^{\perp}\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}^2]^2}{\mathrm{E}[\mathcal{P}_{\boldsymbol{X}}^{\perp}\{RS + (1-R)\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}]^2}.
$$

To show that the proposed test is more powerful than the imputation score test without stratification or the B-spline terms, we consider two sets of linear predictors $\boldsymbol{Z}_1(\widehat{\boldsymbol{\gamma}})$ and

$\boldsymbol{Z}_2(\widehat{\boldsymbol{\gamma}})$, where $\boldsymbol{Z}_1(\boldsymbol{\gamma})$ is contained in $\boldsymbol{Z}_2(\boldsymbol{\gamma})$. Let $\widetilde{S}_1$ and $\widetilde{S}_2$ be the imputed values of S using $\boldsymbol{Z}_1(\boldsymbol{\gamma}_0)$ and $\boldsymbol{Z}_2(\boldsymbol{\gamma}_0)$, respectively. The score test with $\boldsymbol{Z}_2(\widehat{\boldsymbol{\gamma}})$ in the imputation model is asymptotically more powerful if and only if

$$\frac{[p_R\mathrm{E}(\mathcal{P}_{\boldsymbol{X}}^{\perp}S)^2+(1-p_R)\mathrm{E}(\mathcal{P}_{\boldsymbol{X}}^{\perp}\widetilde{S}_2)^2]^2}{\mathrm{E}[\mathcal{P}_{\boldsymbol{X}}^{\perp}\{RS+(1-R)\widetilde{S}_2\}]^2} \geq \frac{[p_R\mathrm{E}(\mathcal{P}_{\boldsymbol{X}}^{\perp}S)^2+(1-p_R)\mathrm{E}(\mathcal{P}_{\boldsymbol{X}}^{\perp}\widetilde{S}_1)^2]^2}{\mathrm{E}[\mathcal{P}_{\boldsymbol{X}}^{\perp}\{RS+(1-R)\widetilde{S}_1\}]^2}. \tag{A.6}$$

After some algebraic manipulation, the denominator on the left side of (A.6) can be expressed as

$$p_R\mathrm{E}(\mathcal{P}_{\boldsymbol{X}}^{\perp}S)^2+(1-p_R)\mathrm{E}(\mathcal{P}_{\boldsymbol{X}}^{\perp}\widetilde{S}_2)^2+p_R(1-p_R)\{\mathrm{E}(S\boldsymbol{X}^{\mathrm{T}})-\mathrm{E}(\widetilde{S}_2\boldsymbol{X}^{\mathrm{T}})\}\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\{\mathrm{E}(S\boldsymbol{X})-\mathrm{E}(\widetilde{S}_2\boldsymbol{X})\}.$$

Compared to $\widetilde{S}_1$, $\widetilde{S}_2$ is the projection of S onto a larger linear space. Thus,

$$\begin{aligned} & \{\mathrm{E}(S\boldsymbol{X}^{\mathrm{T}})-\mathrm{E}(\widetilde{S}_2\boldsymbol{X}^{\mathrm{T}})\}\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\{\mathrm{E}(S\boldsymbol{X})-\mathrm{E}(\widetilde{S}_2\boldsymbol{X})\} \\ \leq\ & \{\mathrm{E}(S\boldsymbol{X}^{\mathrm{T}})-\mathrm{E}(\widetilde{S}_1\boldsymbol{X}^{\mathrm{T}})\}\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\{\mathrm{E}(S\boldsymbol{X})-\mathrm{E}(\widetilde{S}_1\boldsymbol{X})\}, \end{aligned}$$

and $\mathrm{E}(\mathcal{P}_{\boldsymbol{X}}^{\perp}\widetilde{S}_2)^2 \geq \mathrm{E}(\mathcal{P}_{\boldsymbol{X}}^{\perp}\widetilde{S}_1)^2$. It follows that (A.6) holds, and the test with a larger set of covariates in the imputation is more powerful.

REFERENCES

Arend, R. C., Londoño-Joshi, A. I., Straughn, J. M., and Buchsbaum, D. J. (2013), "The Wnt/β-Catenin Pathway in Ovarian Cancer: A Review," *Gynecologic Oncology*, 131, 772–779.

Auer, P. L., Johnsen, J. M., Johnson, A. D., Logsdon, B. A., Lange, L. A., Nalls, M. A., Zhang, G., Franceschini, N., Fox, K., Lange, E. M. et al. (2012), "Imputation of Exome Sequence Variants Into Population-Based Samples and Blood-Cell-Trait-Associated Loci

in African Americans: NHLBI GO Exome Sequencing Project," *The American Journal of Human Genetics*, 91, 794–808.

Bignotti, E., Tassi, R. A., Calza, S., Ravaggi, A., Bandiera, E., Rossi, E., Donzelli, C., Pasinetti, B., Pecorelli, S., and Santin, A. D. (2007), "Gene Expression Profile of Ovarian Serous Papillary Carcinomas: Identification of Metastasis-associated Genes," *American Journal of Obstetrics & Gynecology*, 196, 245–e1.

De Boor, C. (1978), *A Practical Guide to Splines*, New York: Springer.

de Cristofaro, T., Di Palma, T., Soriano, A. A., Monticelli, A., Affinito, O., Cocozza, S., and Zannini, M. (2016), "Candidate Genes and Pathways Downstream of PAX8 Involved in Ovarian High-grade Serous Carcinoma," *Oncotarget*, 7, 41929.

Denkert, C., Schmitt, W. D., Berger, S., Reles, A., Pest, S., Siegert, A., Lichtenegger, W., Dietel, M., and Hauptmann, S. (2002), "Expression of Mitogen-activated Protein Kinase Phosphatase-1 (MKP-1) in Primary Human Ovarian Carcinoma," *International Journal of Cancer*, 102, 507–513.

Derkach, A., Lawless, J. F., and Sun, L. (2015), "Score Tests for Association Under Response-Dependent Sampling Designs for Expensive Covariates," *Biometrika*, 102, 988–994.

Hu, Y. J., Li, Y., Auer, P. L., and Lin, D. Y. (2015), "Integrative Analysis of Sequencing and Array Genotype Data for Discovering Disease Associations With Rare Mutations," *Proceedings of the National Academy of Sciences*, 112, 1019–1024.

Huang, Y. T., VanderWeele, T. J., and Lin, X. (2014), "Joint Analysis of SNP and Gene Expression Data in Genetic Association Studies of Complex Diseases," *The Annals of Applied Statistics*, 8, 352–376.

Kim, H., Golub, G. H., and Park, H. (2005), "Missing Value Estimation for DNA Microarray Gene Expression Data: Local Least Squares Imputation," *Bioinformatics*, 21, 187–198.

Lawless, J. (2016), "Two-phase Outcome-Dependent Studies for Failure Times and Testing for Effects of Expensive Covariates," *Lifetime Data Analysis* [online], DOI:10.1007/s10985-016-9386-8. Available at *https://link.springer.com/journal/10985*.

Li, Y., Willer, C. J., Ding, J., Scheet, P., and Abecasis, G. R. (2010), "MaCH: Using Sequence and Genotype Data to Estimate Haplotypes and Unobserved Genotypes," *Genetic Epidemiology*, 34, 816–834.

Lin, D. Y., Zeng, D., and Tang, Z. Z. (2013), "Quantitative Trait Analysis in Sequencing Studies Under Trait-Dependent Sampling," *Proceedings of the National Academy of Sciences*, 110, 12247–12252.

Little, R. J. (1992), "Regression With Missing X's: A Review," *Journal of the American Statistical Association*, 87, 1227–1237.

Mougeot, J.-L. C., Bahrani-Mostafavi, Z., Vachris, J. C., McKinney, K. Q., Gurlov, S., Zhang, J., Naumann, R. W., Higgins, R. V., and Hall, J. B. (2006), "Gene Expression Profiling of Ovarian Tissues for Determination of Molecular Pathways Reflective of Tumorigenesis," *Journal of molecular biology*, 358, 310–329.

Prentice, R. L., and Pyke, R. (1979), "Logistic Disease Incidence Models and Case-Control Studies," *Biometrika*, 66, 403–411.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, Wiley: New York.

Schumaker, L. (2007), *Spline Functions: Basic Theory*, Cambridge: Cambridge University Press.

Sherman-Baust, C. A., Weeraratna, A. T., Rangel, L. B., Pizer, E. S., Cho, K. R., Schwartz, D. R., Shock, T., and Morin, P. J. (2003), "Remodeling of the Extracellular Matrix Through Overexpression of Collagen VI Contributes to Cisplatin Resistance in Ovarian Cancer Cells," *Cancer Cell*, 3, 377–386.

Tang, Z. Z., and Lin, D. Y. (2013), "MASS: Meta-Analysis of Score Statistics for Sequencing Studies," *Bioinformatics*, 29, 1803–1805.

The Cancer Genome Atlas Research Network (2011), "Integrated Genomic Analyses of Ovarian Carcinoma," *Nature*, 474, 609–615.

Torres-García, W., Zhang, W., Runger, G. C., Johnson, R. H., and Meldrum, D. R. (2009), "Integrative Analysis of Transcriptomic and Proteomic Data of Desulfovibrio Vulgaris: A Non-Linear Model to Predict Abundance of Undetected Proteins," *Bioinformatics*, 25, 1905–1914.

van der Vaart, A. W., and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, New York: Springer.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011), "Rare-variant Association Testing for Sequencing Data With the Sequence Kernel Association Test," *The American Journal of Human Genetics*, 89, 82–93.

Yu, B., Zheng, Y., Alexander, D., Morrison, A. C., Coresh, J., and Boerwinkle, E. (2014), "Genetic Determinants Influencing Human Serum Metabolome Among African Americans," *PLoS Genetics*, 10: e1004212.