Received: 2 March 2019 | Revised: 9 March 2020 | Accepted: 24 March 2020

DOI: 10.1111/biom.13270

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Retrospective versus prospective score tests for genetic association with case-control data

Yukun Liu[1] | **Pengfei Li**[2] | **Lei Song**[3,4] | **Kai Yu**[3] | **Jing Qin**[5]

[1]KLATASDS-MOE, School of Statistics, East China Normal University, Shanghai, China

[2]Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada

[3]National Cancer Institute, National Institutes of Health, Bethesda, Maryland

[4]Cancer Genomics Research Laboratory, Leidos Biomedical Research, Inc., Frederick, Maryland

[5]National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland

Correspondence
Pengfei Li, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada N2L 3G1.
Email: pengfei.li@uwaterloo.ca

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 11771144, 11971300, 11871287; Natural Sciences and Engineering Research Council of Canada, Grant/Award Number: RGPIN-2020-04964; Chinese Ministry of Education 111 Project, Grant/Award Number: B14019; State Key Program of the National Natural Science Foundation of China, Grant/Award Number: 71931004; Natural Science Foundation of Shanghai, Grant/Award Numbers: 17ZR1409000, 19ZR1420900

Abstract

Since the seminal work of Prentice and Pyke, the prospective logistic likelihood has become the standard method of analysis for retrospectively collected case-control data, in particular for testing the association between a single genetic marker and a disease outcome in genetic case-control studies. In the study of multiple genetic markers with relatively small effects, especially those with rare variants, various aggregated approaches based on the same prospective likelihood have been developed to integrate subtle association evidence among all the markers considered. Many of the commonly used tests are derived from the prospective likelihood under a common-random-effect assumption, which assumes a common random effect for all subjects. We develop the locally most powerful aggregation test based on the retrospective likelihood under an independent-random-effect assumption, which allows the genetic effect to vary among subjects. In contrast to the fact that disease prevalence information cannot be used to improve efficiency for the estimation of odds ratio parameters in logistic regression models, we show that it can be utilized to enhance the testing power in genetic association studies. Extensive simulations demonstrate the advantages of the proposed method over the existing ones. A real genome-wide association study is analyzed for illustration.

KEYWORDS

genetic association study, logistic regression model, prospective likelihood, random effect, retrospective likelihood, score test

1 | INTRODUCTION

In genome-wide association studies (GWASs) of relatively rare disease outcomes, such as rare cancers, the case and control study is a commonly used design because of its convenience and cost effectiveness. In a case-control design, a fixed number of cases and controls is used to gather covariate information. Given this information, the most popular model for the disease status is the logistic regression model. Since the seminal paper by Prentice and Pyke (1979), it has been well known that one may use the prospective logistic likelihood to make inference for the underlying odds ratio parameters even if the data are collected retrospectively. In general, the disease prevalence cannot be estimated based on case and control data. Even if available it cannot be used to improve the estimation of the odds ratio parameters. Many statistical genetics papers derive testing statistics based on the prospective logistic likelihood and then apply them without any justification to case and control data. This strategy usually works because the maximum likelihood estimators for the odds ratio

under both likelihoods are the same even though those for the intercept may differ from each other, as shown by Prentice and Pyke (1979). In this paper, however, we show that in some applications the score statistic derived from the retrospective likelihood has a larger power than that derived from the prospective likelihood.

An initial step in the GWAS of a disease is to test whether a specific group of genetic markers have a simultaneous effect on the disease. Denote the disease status as $D = 0$ (no disease) or 1 (disease). Let X be a d-variate vector of the clinical covariates and Y a q-variate vector representing measures on the set of genetic markers considered, such as correlated markers within a candidate region or a gene. The goal of this paper is to study the joint association test of Y with the disease status D after adjusting for the nongenetic covariate X. The logistic regression model for the disease status is $\text{pr}(D = 1|\mathbf{x}, \mathbf{y}) = \pi(\alpha_p + \mathbf{x}^\top\boldsymbol{\beta} + \mathbf{y}^\top\boldsymbol{\gamma})$, where $\pi(t) = e^t/(1 + e^t)$ is the logistic link function. However, the linear function $\mathbf{y}^\top\boldsymbol{\gamma}$ of $\mathbf{y}$ in the model may not be general enough to capture more realistic scenarios, where different genetic markers convey nonuniform risk levels (magnitude and/or direction). A popular approach is to utilize a random-effect model:

$$\text{pr}(D = 1|\mathbf{x}, \mathbf{y}, \mathbf{v}) = \pi(\alpha_p + \mathbf{x}^\top\boldsymbol{\beta} + \mathbf{y}^\top\boldsymbol{\gamma} + \mathbf{y}^\top\mathbf{v} \cdot \sqrt{\theta}), \quad (1)$$

where $\boldsymbol{\gamma}$ and $\mathbf{v}$ denote, respectively, the fixed and random-variant effects. If $h(\mathbf{v})$ denotes the density function of $\mathbf{v}$, we have a marginal probability $\text{pr}(D = 1|\mathbf{x}, \mathbf{y}) = \int \pi(\alpha_p + \mathbf{x}^\top\boldsymbol{\beta} + \mathbf{y}^\top\boldsymbol{\gamma} + \mathbf{y}^\top\mathbf{v} \cdot \sqrt{\theta})h(\mathbf{v})d\mathbf{v}$, which is no longer a logistic regression model. If we specify the random-effect density $h(\mathbf{v})$, then it is possible to derive the likelihood ratio statistic by numerical integration. However, this integration is a formidable task, especially when we examine thousands of genes or regions. The resulting likelihood ratio test would lose power when the random-effect density is misspecified.

Under Model (1), testing the nonexistence of genetic effects is equivalent to testing $H_0 : \boldsymbol{\gamma} = 0 \& \theta = 0$. The score test is popular in the statistical and genetics literature since it is evaluated at the null hypothesis and can effectively avoid the need to specify the form of $h(\mathbf{v})$. Many score tests have been developed under various model assumptions. For example, the burden and adaptive burden tests are designed under $\theta = 0$. Assuming $\boldsymbol{\gamma} = 0$, Wu *et al.* (2011) proposed a sequence kernel association test (SKAT), which Lee *et al.* (2012) extended to SKAT-O by integrating the SKAT and burden tests into a single test. Without assuming either $\theta = 0$ or $\boldsymbol{\gamma} = 0$, Sun *et al.* (2013) proposed a mixed-effects score test (MiST) by combining information from the fixed and random-variant effects.

The SKAT, SKAT-O, and MiST tests are all built on the assumption that observations from individuals share the same random-effect $\mathbf{v}$ and are independent given $\mathbf{v}$. In other words, not all the subjects are independent. This assumption is fundamentally different from the conventional random-effect model assumption (called the independent-random-effect assumption hereafter), which treats observations as independent of each other unless they come from the same individual or cluster (Verbeke and Lesaffre, 1996; Wang, 1998; Ke and Wang, 2001; Jiang, 2007). This motivates us to develop score tests for the genetic effect based on case-control data under the independent random-effect model. Specifically, suppose that $\{(\mathbf{x}_i, \mathbf{y}_i, D_i, \mathbf{v}_i) : i = 1, 2, \ldots, n\}$ are independent and identically distributed (iid) from Model (1). We observe not the random-effects $\mathbf{v}_i$ but $\{(\mathbf{x}_i, \mathbf{y}_i, D_i) : i = 1, 2, \ldots, n\}$. The observations are independent since the random effect is not shared among them. For convenience of presentation, we assume that $\boldsymbol{\gamma} = \gamma\mathbf{1}$ and the components of the $\mathbf{v}_i$s are uncorrelated with mean zero and variance 2. This variance affects only the score function with the random effect by a multiplier, and does not affect our resulting standardized score test statistic.

We will systematically investigate different score test statistics based on prospectively and retrospectively collected data under the independent random-effect model. In Section 2, we highlight the connections and differences for testing the nonexistence of a random effect between the case-control and prospective designs. Section 3 establishes the asymptotic normality of the score statistics derived in Section 2. In Section 4, we construct several synthetic score tests for the overall genetic effect. Section 5 presents a simulation study, and Section 6 discusses our real-data analysis. We provide some discussion in Section 7. Interestingly, we have found that the independent random-effect model plays a more important role than the retrospective likelihood in the efficiency gain of the proposed tests over the existing tests. In addition, knowledge on the disease prevalence can indeed enhance power when testing for the existence of a random effect. For brevity, the technical details and additional simulation results are given in the Supporting Information.

2 | RETROSPECTIVE AND PROSPECTIVE SCORE STATISTICS

2.1 | Retrospective likelihood

Let $\boldsymbol{\varphi} = (\alpha_p, \boldsymbol{\beta}, \gamma, \theta)^\top$, $f(\mathbf{x}, \mathbf{y})$ be the joint density function of $(\mathbf{x}, \mathbf{y})$, and

$$g(\mathbf{x}, \mathbf{y}, \boldsymbol{\varphi}) = \text{pr}(D = 1|\mathbf{x}, \mathbf{y}) = \int \pi(\alpha_p + \mathbf{x}^\top\boldsymbol{\beta} + \mathbf{y}^\top\mathbf{1}\gamma + \mathbf{y}^\top\mathbf{v} \cdot \sqrt{\theta})h(\mathbf{v})d\mathbf{v},$$

where $h(\mathbf{v})$ is the density function of $\mathbf{v}$. Using Bayes' formula, we find that the densities of the covariates in the cases and

controls are

$$\begin{aligned} f_1(\mathbf{x},\mathbf{y},\boldsymbol{\varphi}) &:= \mathrm{pr}(\mathbf{x},\mathbf{y}|D=1) \\ &= g(\mathbf{x},\mathbf{y},\boldsymbol{\varphi})f(\mathbf{x},\mathbf{y}) \Big/ \int g(\mathbf{x},\mathbf{y},\boldsymbol{\varphi})f(\mathbf{x},\mathbf{y})d\mathbf{x}d\mathbf{y}, \\ f_0(\mathbf{x},\mathbf{y},\boldsymbol{\varphi}) &:= \mathrm{pr}(\mathbf{x},\mathbf{y}|D=0) \\ &= \{1-g(\mathbf{x},\mathbf{y},\boldsymbol{\varphi})\}f(\mathbf{x},\mathbf{y}) \Big/ \\ &\quad \times \{1-\int g(\mathbf{x},\mathbf{y},\boldsymbol{\varphi})f(\mathbf{x},\mathbf{y})d\mathbf{x}d\mathbf{y}\}, \end{aligned}$$

respectively. Without loss of generality, we assume that the first n_0 of $\{(\mathbf{x}_i,\mathbf{y}_i,D_i) : i=1,2,\ldots,n\}$ are controls and the last $n_1 = n - n_0$ are cases. The retrospective likelihood for case and control data is

$$L_{\mathrm{retr}}(\boldsymbol{\varphi}) = \prod_{i=1}^{n}[\{f_1(\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\varphi})\}^{D_i}\{f_0(\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\varphi})\}^{1-D_i}]. \quad (2)$$

It is necessary to make a direct comparison of the retrospective likelihood in (2) and Lin's (1997) prospective likelihood (defined below), since the former is the foundation of this paper while the latter is the infrastructure in variance-component score tests such as SKAT, SKAT-O, and MiST.

In the derivation of (2), we have assumed that the random-effects $\mathbf{v}_i$ for different individuals are independent of each other. In contrast, Lin's (1997) likelihood is derived assuming common random effects and prospective data. Under these assumptions, Lin's (1997) likelihood function is

$$L(\boldsymbol{\varphi}) = \int \prod_{i=1}^{n}\{\pi_i(\mathbf{v},\boldsymbol{\varphi})\}^{D_i}\{1-\pi_i(\mathbf{v},\boldsymbol{\varphi})\}^{1-D_i}h(\mathbf{v})d\mathbf{v}, \quad (3)$$

where $\pi_i(\mathbf{v},\boldsymbol{\varphi}) = \pi(\alpha_p + \mathbf{x}_i^\top\boldsymbol{\beta} + \mathbf{y}_i^\top\mathbf{1}\gamma + \mathbf{y}_i^\top\mathbf{v}\cdot\sqrt{\theta})$. In this model, it seems impossible to find a valid variance estimator for any point estimators since no replicates from the random-effect $\mathbf{v}$ are available. In contrast, under the independent random-effect model, the likelihood is

$$L_C(\boldsymbol{\varphi}) = \prod_{i=1}^{n}\int \{\pi_i(\mathbf{v},\boldsymbol{\varphi})\}^{D_i}\{1-\pi_i(\mathbf{v},\boldsymbol{\varphi})\}^{1-D_i}h(\mathbf{v})d\mathbf{v}.$$

Compared with $L(\boldsymbol{\varphi})$, here the order of integration and product is changed.

In summary, L_{retr} and L_C are the retrospective and prospective likelihoods respectively under the independent-random-effect assumption, and L is the prospective likelihood under the common-random-effect assumption. The retrospective likelihood under the common-random-effect assumption is quite complicated and we defer its detailed derivation to the Supporting Information.

For a test of $H_0 : \gamma = 0 \;\&\; \theta = 0$, it is widely accepted that score tests are preferable to likelihood ratio tests because they avoid using the form of $h(\mathbf{v})$. The score tests derived from (3) with respect to γ and θ are usually correlated. Therefore, it is not straightforward to combine the two score statistics to achieve satisfactory power under different types of alternatives. To overcome this obstacle, Sun *et al.* (2013) derived a new score statistic with respect to θ without necessarily requiring $\gamma = 0$. Their approach may have power loss since their derivations are based on a prospective likelihood while the available data are retrospectively collected. In this paper, we find the retrospective-likelihood–based score tests directly and explore their relationship with Sun *et al.*'s (2013) score tests.

2.2 | Retrospective score statistic for θ

Differentiating the retrospective log-likelihood $\ell_{\mathrm{retr}}(\boldsymbol{\varphi}) = \log\{L_{\mathrm{retr}}(\boldsymbol{\varphi})\}$ with respect to θ, we find that the score is

$$\begin{aligned} \frac{\partial \ell_{\mathrm{retr}}}{\partial\theta}\Big|_{\theta=0} &= \sum_{i=1}^{n}\left\{D_i - \pi\left(\alpha_p + \boldsymbol{\beta}^\top\mathbf{x}_i + \mathbf{y}_i^\top\mathbf{1}\gamma\right)\right\} \\ &\quad \times \left\{1 - 2\pi\left(\alpha_p + \boldsymbol{\beta}^\top\mathbf{x}_i + \mathbf{y}_i^\top\mathbf{1}\gamma\right)\right\}\mathbf{y}_i^\top\mathbf{y}_i \\ &\quad + n_0\int\left\{1 - 2\pi\left(\alpha_p + \boldsymbol{\beta}^\top\mathbf{x} + \mathbf{y}^\top\mathbf{1}\gamma\right)\right\} \\ &\quad \times \pi(\alpha_p + \boldsymbol{\beta}^\top\mathbf{x} + \mathbf{y}^\top\mathbf{1}\gamma)\mathbf{y}^\top\mathbf{y}f_0(\mathbf{x},\mathbf{y})d\mathbf{x}d\mathbf{y} \\ &\quad - n_1\int\left\{1 - 2\pi\left(\alpha_p + \boldsymbol{\beta}^\top\mathbf{x} + \mathbf{y}^\top\mathbf{1}\gamma\right)\right\} \\ &\quad \times \left\{1 - \pi\left(\alpha_p + \boldsymbol{\beta}^\top\mathbf{x} + \mathbf{y}^\top\mathbf{1}\gamma\right)\right\}\mathbf{y}^\top\mathbf{y}f_1(\mathbf{x},\mathbf{y})d\mathbf{x}d\mathbf{y}. \end{aligned}$$

To implement the score test statistic, we need to estimate the unknown parameters $(\alpha_p,\boldsymbol{\beta},\gamma)$ under $\theta = 0$. Note that the density functions of the cases and controls satisfy the density ratio model (Anderson, 1979; Qin and Zhang, 1997; Qin, 2017),

$$f_1(\mathbf{x},\mathbf{y},\boldsymbol{\varphi}) = \exp(\alpha_r + \boldsymbol{\beta}^\top\mathbf{x} + \mathbf{y}^\top\mathbf{1}\gamma)f_0(\mathbf{x},\mathbf{y},\boldsymbol{\varphi}),$$

where $\alpha_r = \alpha_p + \log\{(1-p)/p\}$ and $p = \mathrm{pr}(D=1)$ denotes the prevalence of the disease of interest. To simplify our notation, we use $\boldsymbol{\beta}$ and αs, such as α_r and α_p, to denote both argument variables and their true values; the meaning is clear from the context. Even if α_r is known, the parameter $\alpha_p = \alpha_r - \log\{(1-p)/p\}$ is generally unknown because the prevalence p is unknown. For the time being, we assume that the prevalence p is known, so the estimation of α_p is equivalent to the estimation of α_r.

Following Qin and Zhang (1997), we estimate $(\alpha_r,\boldsymbol{\beta},\gamma)$ by the maximizer $(\tilde{\alpha}_r,\tilde{\boldsymbol{\beta}},\tilde{\gamma})$ of

$$\begin{aligned} \ell_{e,1}(\alpha_r,\boldsymbol{\beta},\gamma) &= \sum_{i=1}^{n} D_i\left(\alpha_r + \boldsymbol{\beta}^\top\mathbf{x}_i + \mathbf{y}_i^\top\mathbf{1}\gamma\right) - \sum_{i=1}^{n}\log \\ &\quad \times \left\{1 + (n_1/n_0)\exp\left(\alpha_r + \boldsymbol{\beta}^\top\mathbf{x}_i + \mathbf{y}_i^\top\mathbf{1}\gamma\right)\right\}. \end{aligned}$$

Let $F_0(\mathbf{x},\mathbf{y})$ and $F_1(\mathbf{x},\mathbf{y})$ be the true distribution functions corresponding to $f_0(\mathbf{x},\mathbf{y},\boldsymbol{\varphi})$ and $f_1(\mathbf{x},\mathbf{y},\boldsymbol{\varphi})$, respectively. Under $\theta = 0$, the maximum empirical likelihood estimators of $F_0(\mathbf{x},\mathbf{y})$ and $F_1(\mathbf{x},\mathbf{y})$ are

$$\tilde{F}_0(\mathbf{x},\mathbf{y}) = \frac{1}{n_0}\sum_{i=1}^{n}\left\{1-\pi\left(\tilde{\alpha}+\tilde{\boldsymbol{\beta}}^{\top}\mathbf{x}_i+\mathbf{y}_i^{\top}\mathbf{1}\tilde{\gamma}\right)\right\}I(\mathbf{x}_i\le\mathbf{x},\mathbf{y}_i\le\mathbf{y})$$

and

$$\tilde{F}_1(\mathbf{x},\mathbf{y}) = n_1^{-1}\sum_{i=1}^{n}\pi\left(\tilde{\alpha}+\tilde{\boldsymbol{\beta}}^{\top}\mathbf{x}_i+\mathbf{y}_i^{\top}\mathbf{1}\tilde{\gamma}\right)I(\mathbf{x}_i\le\mathbf{x},\mathbf{y}_i\le\mathbf{y}),$$

where $\tilde{\alpha}=\tilde{\alpha}_r+\log(n_1/n_0)$, $I(\cdot)$ is the indicator function, and the inequality $\mathbf{x}_i\le\mathbf{x}$ holds elementwise.

Putting these estimators into $\partial\ell_{\text{retr}}/\partial\theta|_{\theta=0}$ leads to

$$U_1(\alpha_p) = \sum_{i=1}^{n}\left\{D_i-\pi\left(\tilde{\alpha}+\tilde{\boldsymbol{\beta}}^{\top}\mathbf{x}_i+\mathbf{y}_i^{\top}\mathbf{1}\tilde{\gamma}\right)\right\} \times\left\{1-2\pi\left(\alpha_p+\tilde{\boldsymbol{\beta}}^{\top}\mathbf{x}_i+\mathbf{y}_i^{\top}\mathbf{1}\tilde{\gamma}\right)\right\}\mathbf{y}_i^{\top}\mathbf{y}_i, \quad (4)$$

which is the retrospective score statistic with respect to θ if α_p is known. We will discuss later the case where α_p is unknown.

2.3 | Retrospective score statistic for γ

Since $\partial g(\mathbf{x},\mathbf{y},\boldsymbol{\varphi})/\partial\gamma|_{\theta=0,\gamma=0} = \pi(\alpha_p+\mathbf{x}^{\top}\boldsymbol{\beta})\{1-\pi(\alpha_p+\mathbf{x}^{\top}\boldsymbol{\beta})\}\mathbf{y}^{\top}\mathbf{1}$, a similar derivation to that for the score statistic with respect to θ gives the retrospective score with respect to γ:

$$\left.\frac{\partial\ell_{\text{retr}}}{\partial\gamma}\right|_{\theta=0,\gamma=0} = \sum_{i=1}^{n}\left\{D_i-\pi\left(\alpha_p+\mathbf{x}_i^{\top}\boldsymbol{\beta}\right)\right\}\mathbf{y}_i^{\top}\mathbf{1} + n_0\int\pi(\alpha_p+\mathbf{x}^{\top}\boldsymbol{\beta})\mathbf{y}^{\top}\mathbf{1}f_1(\mathbf{x},\mathbf{y})d\mathbf{x}d\mathbf{y}-n_1 \times\int\left\{1-\pi(\alpha_p+\mathbf{x}^{\top}\boldsymbol{\beta})\right\}\mathbf{y}^{\top}\mathbf{1}f_0(\mathbf{x},\mathbf{y})d\mathbf{x}d\mathbf{y}.$$

To estimate the unknown parameters $(\alpha_p,\boldsymbol{\beta})$ under H_0, recall that the density functions of the cases and controls are linked by $f_1(\mathbf{x},\mathbf{y},\boldsymbol{\varphi}) = \exp(\alpha_r+\boldsymbol{\beta}^{\top}\mathbf{x})f_0(\mathbf{x},\mathbf{y},\boldsymbol{\varphi})$. Following Qin and Zhang (1997), we estimate $(\alpha_r,\boldsymbol{\beta})$ by the maximizer $(\hat{\alpha}_r,\hat{\boldsymbol{\beta}})$ of

$$\ell_1(\alpha_r,\boldsymbol{\beta}) = \sum_{i=1}^{n}D_i(\alpha_r+\boldsymbol{\beta}^{\top}\mathbf{x}_i) - \sum_{i=1}^{n}\log\{1+(n_1/n_0)\exp(\alpha_r+\boldsymbol{\beta}^{\top}\mathbf{x}_i)\}.$$

Under H_0, the maximum empirical likelihood estimators of $F_0(\mathbf{x},\mathbf{y})$ and $F_1(\mathbf{x},\mathbf{y})$ are

$$\hat{F}_0(\mathbf{x},\mathbf{y}) = n_0^{-1}\sum_{i=1}^{n}\{1-\pi(\hat{\alpha}+\hat{\boldsymbol{\beta}}^{\top}\mathbf{x}_i)\}I(\mathbf{x}_i\le\mathbf{x},\mathbf{y}_i\le\mathbf{y})$$

and

$$\hat{F}_1(\mathbf{x},\mathbf{y}) = n_1^{-1}\sum_{i=1}^{n}\pi(\hat{\alpha}+\hat{\boldsymbol{\beta}}^{\top}\mathbf{x}_i)I(\mathbf{x}_i\le\mathbf{x},\mathbf{y}_i\le\mathbf{y}),$$

where $\hat{\alpha}=\hat{\alpha}_r+\log(n_1/n_0)$.

Putting $\hat{\alpha}$, $\hat{\boldsymbol{\beta}}$, $\hat{F}_0(\mathbf{x},\mathbf{y})$, and $\hat{F}_1(\mathbf{x},\mathbf{y})$ into $\partial\ell_{\text{retr}}/\partial\gamma|_{\theta=0,\gamma=0}$, we have the retrospective score statistic with respect to γ:

$$U_2 = \sum_{i=1}^{n}\{D_i-\pi(\hat{\alpha}+\hat{\boldsymbol{\beta}}^{\top}\mathbf{x}_i)\}\mathbf{y}_i^{\top}\mathbf{1}, \quad (5)$$

which is independent of α_p.

2.4 | Prospective score statistic with θ

If we treat the case-control data as if they were collected prospectively, i.e., $\{(D_i,\mathbf{x}_i,\mathbf{y}_i): i=1,2,\ldots,n\}$ are iid random elements, the resulting prospective log-likelihood is

$$\ell_{\text{pros}} = \sum_{i=1}^{n}[D_i\log\{g(\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\varphi})\}+(1-D_i) \times\log\{1-g(\mathbf{x}_i,\mathbf{y}_i,\boldsymbol{\varphi})\}]+\sum_{i=1}^{n}\log\{f(\mathbf{x}_i,\mathbf{y}_i)\}.$$

The prospective score with respect to θ without restricting γ to 0 is

$$\left.\frac{\partial\ell_{\text{pros}}}{\partial\theta}\right|_{\theta=0} = \sum_{i=1}^{n}\left\{D_i-\pi\left(\alpha_p+\boldsymbol{\beta}^{\top}\mathbf{x}_i+\mathbf{y}_i^{\top}\mathbf{1}\gamma\right)\right\} \times\left\{1-2\pi\left(\alpha_p+\boldsymbol{\beta}^{\top}\mathbf{x}_i+\mathbf{y}_i^{\top}\mathbf{1}\gamma\right)\right\}\mathbf{y}_i^{\top}\mathbf{y}_i. \quad (6)$$

The unknown parameters α_p, $\boldsymbol{\beta}$, and γ are estimated by the maximum prospective likelihood estimator under $\theta=0$. The prospective likelihood ℓ_{pros} under $\theta=0$ reduces to (up to a quantity independent of the parameters)

$$\ell_{e,2}(\alpha_p,\boldsymbol{\beta},\gamma) = \sum_{i=1}^{n}D_i\left(\alpha_p+\boldsymbol{\beta}^{\top}\mathbf{x}_i+\mathbf{y}_i^{\top}\mathbf{1}\gamma\right) - \sum_{i=1}^{n}\log\left\{1+\exp\left(\alpha_p+\boldsymbol{\beta}^{\top}\mathbf{x}_i+\mathbf{y}_i^{\top}\mathbf{1}\gamma\right)\right\}.$$

Hence, the maximum likelihood estimator of $(\alpha_p,\boldsymbol{\beta},\gamma)$ is $(\check{\alpha}_p,\check{\boldsymbol{\beta}},\check{\gamma}) = \arg\max_{\alpha_p,\boldsymbol{\beta},\gamma}\ell_{e,2}(\alpha_p,\boldsymbol{\beta},\gamma)$.

Denoting $\alpha = \alpha_r + \log(n_1/n_0) = \alpha_p + \log\{(1-p)/p\} + \log(n_1/n_0)$, we have $\ell_{e,1}(\alpha_r, \boldsymbol{\beta}, \gamma) = \ell_{e,2}(\alpha, \boldsymbol{\beta}, \gamma) - n_1 \log(n_1/n_0)$. Since $\tilde{\alpha} = \tilde{\alpha}_r + \log(n_1/n_0)$, it follows that

$$(\tilde{\alpha}, \widetilde{\boldsymbol{\beta}}, \tilde{\gamma}) = \arg\max_{\alpha,\boldsymbol{\beta},\gamma} \ell_{e,2}(\alpha, \boldsymbol{\beta}, \gamma)$$

$$= \arg\max_{\alpha_p,\boldsymbol{\beta},\gamma} \ell_{e,2}(\alpha_p, \boldsymbol{\beta}, \gamma) = (\breve{\alpha}_p, \breve{\boldsymbol{\beta}}, \breve{\gamma}).$$

Replacing $(\alpha_p, \boldsymbol{\beta}, \gamma)$ by $(\breve{\alpha}_p, \breve{\boldsymbol{\beta}}, \breve{\gamma}) = (\tilde{\alpha}, \widetilde{\boldsymbol{\beta}}, \tilde{\gamma})$ in (6), we obtain the prospective score statistic with respect to θ:

$$\sum_{i=1}^{n} \left\{ D_i - \pi\left(\tilde{\alpha} + \widetilde{\boldsymbol{\beta}}^\top \mathbf{x}_i + \mathbf{y}_i^\top \mathbf{1}\tilde{\gamma}\right) \right\} \times \left\{ 1 - 2\pi\left(\tilde{\alpha} + \widetilde{\boldsymbol{\beta}}^\top \mathbf{x}_i + \mathbf{y}_i^\top \mathbf{1}\tilde{\gamma}\right) \right\} \mathbf{y}_i^\top \mathbf{y}_i = U_1(\tilde{\alpha}), \quad (7)$$

where $U_1(\cdot)$ is defined in (4). That is, the only difference between the prospective score statistic and the retrospective score statistic with respect to θ is the use of different α values in $\{1 - 2\pi(\alpha + \widetilde{\boldsymbol{\beta}}^\top \mathbf{x}_i + \mathbf{y}_i^\top \mathbf{1}\tilde{\gamma})\}\mathbf{y}_i^\top \mathbf{y}_i$. We will show that this difference can lead to a severe power loss in the hypothesis testing for a genetic effect.

2.5 | Prospective score statistic with respect to γ

By direct calculations, we have the prospective score with respect to γ:

$$\left.\frac{\partial \ell_{\text{pros}}}{\partial \gamma}\right|_{\theta=0,\gamma=0} = \sum_{i=1}^{n} \{ D_i - \pi(\alpha_p + \boldsymbol{\beta}^\top \mathbf{x}_i) \} \mathbf{y}_i^\top \mathbf{1}. \quad (8)$$

We estimate the unknown parameters α_p and $\boldsymbol{\beta}$ by maximizing the prospective likelihood estimator under H_0, which up to a quantity independent of the parameters is

$$\ell_2(\alpha_p, \boldsymbol{\beta}) = \sum_{i=1}^{n} D_i(\alpha_p + \boldsymbol{\beta}^\top \mathbf{x}_i) - \sum_{i=1}^{n} \log\{1 + \exp(\alpha_p + \boldsymbol{\beta}^\top \mathbf{x}_i)\}. \quad (9)$$

Hence, the maximum likelihood of $(\alpha_p, \boldsymbol{\beta})$ is given by $(\breve{\alpha}_p, \breve{\boldsymbol{\beta}}) = \arg\max_{\alpha_p,\boldsymbol{\beta}} \ell_2(\alpha_p, \boldsymbol{\beta})$.

Note that $\ell_1(\alpha_r, \boldsymbol{\beta}) = \ell_2(\alpha, \boldsymbol{\beta}) - n_1 \log(n_1/n_0)$ because $\alpha = \alpha_r + \log(n_1/n_0)$. Since $\hat{\alpha} = \hat{\alpha}_r + \log(n_1/n_0)$, it follows that $(\hat{\alpha}, \widehat{\boldsymbol{\beta}}) = \arg\max_{\alpha,\boldsymbol{\beta}} \ell_2(\alpha, \boldsymbol{\beta}) = \arg\max_{\alpha_p,\boldsymbol{\beta}} \ell_2(\alpha_p, \boldsymbol{\beta}) = (\breve{\alpha}_p, \breve{\boldsymbol{\beta}})$, where $\hat{\alpha}, \hat{\alpha}_r$, and $\widehat{\boldsymbol{\beta}}$ are defined in Section 2.3. Replacing $(\alpha_p, \boldsymbol{\beta})$ by $(\breve{\alpha}_p, \breve{\boldsymbol{\beta}}) = (\hat{\alpha}, \widehat{\boldsymbol{\beta}})$ in (8), we find that the prospective score statistic with respect to γ is exactly equal to U_2, which is the retrospective score statistic with respect to γ.

3 | ASYMPTOTICS

This section studies the limiting distributions of the retrospective and prospective score statistics in (4), (5), and (7) for both retrospectively and prospectively collected data.

3.1 | Asymptotic normality

For convenience we assume that n_0/n is a constant as $n \to \infty$, where $n = n_0 + n_1$. Let $\boldsymbol{\xi}_0^\top = (\alpha^\top, \boldsymbol{\beta}^\top)$ and $\boldsymbol{\xi}_0^\top = (\alpha_p^\top, \boldsymbol{\beta}^\top)$, respectively, be the true parameter values for retrospective and prospective data. Denote $\boldsymbol{\xi}_* = (\alpha_*, \boldsymbol{\beta}^\top)^\top$, $\mathbf{z}_i = (1, \mathbf{x}_i^\top)^\top$, $\mathbf{z}_{e,i} = (1, \mathbf{x}_i^\top, \mathbf{y}_i^\top \mathbf{1})^\top$, $\mathbf{z} = (1, \mathbf{x}^\top)^\top$, and $\mathbf{z}_e = (1, \mathbf{x}^\top, \mathbf{y}^\top \mathbf{1})^\top$. Define $C_2(\mathbf{x}, \mathbf{y}) = \mathbf{y}^\top \mathbf{1} - \mathbf{S}_{xy}^\top \mathbf{S}_x^{-1} \mathbf{z}$ and

$$C_1(\mathbf{x}, \mathbf{y}, \alpha_*) = \left\{ 1 - 2\pi\left(\boldsymbol{\xi}_*^\top \mathbf{z}_i\right) \right\} \mathbf{y}^\top \mathbf{y} - \mathbf{H}(\alpha_*) \mathbf{S}_{e,x}^{-1} \mathbf{z}_e,$$

where

$$\mathbf{S}_x = n^{-1} \mathbb{E}\left[\sum_{i=1}^{n} \pi\left(\boldsymbol{\xi}_0^\top \mathbf{z}_i\right) \left\{ 1 - \pi\left(\boldsymbol{\xi}_0^\top \mathbf{z}_i\right) \right\} \mathbf{z}_i \mathbf{z}_i^\top \right],$$

$$\mathbf{S}_{e,x} = n^{-1} \mathbb{E}\left[\sum_{i=1}^{n} \pi\left(\boldsymbol{\xi}_0^\top \mathbf{z}_i\right) \left\{ 1 - \pi\left(\boldsymbol{\xi}_0^\top \mathbf{z}_i\right) \right\} \mathbf{z}_{e,i} \mathbf{z}_{e,i}^\top \right],$$

$$\mathbf{S}_{xy} = n^{-1} \mathbb{E}\left[\sum_{i=1}^{n} \pi\left(\boldsymbol{\xi}_0^\top \mathbf{z}_i\right) \left\{ 1 - \pi\left(\boldsymbol{\xi}_0^\top \mathbf{z}_i\right) \right\} \mathbf{y}_i^\top \mathbf{1} \mathbf{z}_i \right],$$

and $\mathbf{H}(\alpha_*) = n^{-1} \mathbb{E}[\sum_{i=1}^{n} \{1 - \pi(\boldsymbol{\xi}_0^\top \mathbf{z}_i)\} \pi(\boldsymbol{\xi}_0^\top \mathbf{z}_i) \{1 - 2\pi(\boldsymbol{\xi}_*^\top \mathbf{z}_i)\} \mathbf{y}_i^\top \mathbf{y}_i \mathbf{z}_{e,i}^\top]$. We remark that all four quantities are independent of n, and the expectation operator $\mathbb{E}$ has different meanings for retrospective and prospective data.

If the $(D_i, \mathbf{x}_i, \mathbf{y}_i)$s are case-control or retrospective data, we define

$$\sigma_{11}(\alpha_1, \alpha_2) = (n_0/n) \mathbb{E}_0\left[\left\{ \pi\left(\boldsymbol{\xi}_0^\top \mathbf{z}_i\right) \right\}^2 C_1(\mathbf{x}_i, \mathbf{y}_i, \alpha_1) C_1(\mathbf{x}_i, \mathbf{y}_i, \alpha_2) \right] + (n_1/n) \mathbb{E}_1\left[\left\{ 1 - \pi\left(\boldsymbol{\xi}_0^\top \mathbf{z}_i\right) \right\}^2 \times C_1(\mathbf{x}_i, \mathbf{y}_i, \alpha_1) C_1(\mathbf{x}_i, \mathbf{y}_i, \alpha_2) \right]$$

and $\sigma_{22} = (n_0/n) \mathbb{E}_0[\{\pi(\boldsymbol{\xi}_0^\top \mathbf{z}_i)\}^2 C_2^2(\mathbf{x}_i, \mathbf{y}_i)] + (n_1/n) \mathbb{E}_1[\{1 - \pi(\boldsymbol{\xi}_0^\top \mathbf{z}_i)\}^2 C_2^2(\mathbf{x}_i, \mathbf{y}_i)]$, where $\mathbb{E}_0$ ($\mathbb{E}_1$) denotes the expectation operator with respect to $f_0(\mathbf{x}, \mathbf{y}, \boldsymbol{\varphi})$ ($f_1(\mathbf{x}, \mathbf{y}, \boldsymbol{\varphi})$) with $\boldsymbol{\varphi}$ taking its true value. If the $(D_i, \mathbf{x}_i, \mathbf{y}_i)$s are prospective data, we define

$$\sigma_{11}(\alpha_1, \alpha_2) = \mathbb{E}\left[\pi(\boldsymbol{\xi}_0^\top \mathbf{z}_i) \left\{ 1 - \pi\left(\boldsymbol{\xi}_0^\top \mathbf{z}_i\right) \right\} C_1(\mathbf{x}_i, \mathbf{y}_i, \alpha_1) C_1(\mathbf{x}_i, \mathbf{y}_i, \alpha_2) \right]$$

and $\sigma_{22} = \mathbb{E}[\pi(\boldsymbol{\xi}_0^\top \mathbf{z}_i)\{1-\pi(\boldsymbol{\xi}_0^\top \mathbf{z}_i)\}C_2^2(\mathbf{x}_i, \mathbf{y}_i)]$, where $\mathbb{E}$ denotes the expectation operator with respect to $f(\mathbf{x}, \mathbf{y})$.

Theorem 1. *Assume that* $\mathbb{E}(\|\mathbf{X}\|^2) + \mathbb{E}(\|\mathbf{Y}\|^3) < \infty$ *and that* n_1/n *is a constant as* n *goes to infinity. For any* m *constants* $\alpha_{*1}, \ldots, \alpha_{*m}$, *if* $\hat{\alpha}_{*1}, \ldots, \hat{\alpha}_{*m}$ *are their consistent estimators, then as* n *goes to infinity,* $n^{-1/2}(U_1(\hat{\alpha}_{*1}), \ldots, U_1(\hat{\alpha}_{*m}), U_2)^\top$ *converges in distribution to* $N(0_{(m+1)\times 1}, \text{diag}(\boldsymbol{\Sigma}(\alpha_{*1}, \ldots, \alpha_{*m}), \sigma_{22}))$, *where* $\boldsymbol{\Sigma}(\alpha_{*1}, \ldots, \alpha_{*m}) = (\sigma_{11}(\alpha_{*i}, \alpha_{*j}))_{1\le i,j\le m}$.

The proof of Theorem 1 is provided in the Supporting Information. The retrospective and prospective score statistics with respect to θ are $U_1(\alpha_p)$ and $U_1(\tilde{\alpha})$, respectively, where α_p is assumed to be known and $\tilde{\alpha} = \alpha + o_p(1)$ with $\alpha = \alpha_p + \log\{(1-p)n_1/(pn_0)\}$. Theorem 1 indicates that $U_1(\hat{\alpha}_*)$ is asymptotically independent of U_2 for any α_* if $\hat{\alpha}_* = \alpha_* + o_p(1)$. It also implies that both $\sqrt{n}U_1(\alpha_p)$ and $\sqrt{n}U_1(\tilde{\alpha})$ converge in distribution to normal distributions with mean zero, but in general their asymptotic variances $\sigma_{11}(\alpha_p, \alpha_p)$ and $\sigma_{11}(\alpha, \alpha)$ are different. If the proportion n_1/n of the cases in the case-control data is equal to the prevalence p, then $\alpha = \alpha_p$ and the retrospective and prospective score tests and their limiting distributions coincide.

3.2 | Estimation of variance matrix

To apply the retrospective or prospective tests we need consistent estimators of the corresponding asymptotic variances. Since $\widetilde{\boldsymbol{\xi}}$ is a root-n consistent estimator of $\boldsymbol{\xi}_0$ whether the data are retrospective or prospective, natural root-n consistent estimators of $\mathbf{S}_x$, $\mathbf{S}_{e,x}$, $\mathbf{S}_{xy}$, and $\mathbf{H}(\alpha_*)$ are

$$\widetilde{\mathbf{S}}_x = n^{-1}\sum_{i=1}^n \pi(\widetilde{\boldsymbol{\xi}}^\top \mathbf{z}_i)\{1-\pi(\widetilde{\boldsymbol{\xi}}^\top \mathbf{z}_i)\}\mathbf{z}_i\mathbf{z}_i^\top,$$

$$\widetilde{\mathbf{S}}_{e,x} = n^{-1}\sum_{i=1}^n \pi(\widetilde{\boldsymbol{\xi}}^\top \mathbf{z}_i)\{1-\pi(\widetilde{\boldsymbol{\xi}}^\top \mathbf{z}_i)\}\mathbf{z}_{e,i}\mathbf{z}_{e,i}^\top,$$

$$\widetilde{\mathbf{S}}_{xy} = n^{-1}\sum_{i=1}^n \pi(\widetilde{\boldsymbol{\xi}}^\top \mathbf{z}_i)\{1-\pi(\widetilde{\boldsymbol{\xi}}^\top \mathbf{z}_i)\}\mathbf{y}_i^\top \mathbf{1}\mathbf{z}_i,$$

and $\widetilde{\mathbf{H}}(\alpha_*) = n^{-1}\sum_{i=1}^n \{1-\pi(\widetilde{\boldsymbol{\xi}}^\top \mathbf{z}_i)\}\pi(\widetilde{\boldsymbol{\xi}}^\top \mathbf{z}_i)\{1-2\pi(\widetilde{\boldsymbol{\xi}}_*^\top \mathbf{z}_i)\}\mathbf{y}_i^\top \mathbf{y}_i \mathbf{z}_{e,i}^\top$, where $\widetilde{\boldsymbol{\xi}}_* = (\alpha_*, \widetilde{\boldsymbol{\beta}})$. Further, define

$$\hat{\sigma}_{11}(\alpha_{*1}, \alpha_{*2}) = n^{-1}\sum_{i=1}^n \pi(\widetilde{\boldsymbol{\xi}}^\top \mathbf{z}_i)\{1-\pi(\widetilde{\boldsymbol{\xi}}^\top \mathbf{z}_i)\} \times \tilde{C}_1(\mathbf{x}_i, \mathbf{y}_i, \alpha_{*1})\tilde{C}_1(\mathbf{x}_i, \mathbf{y}_i, \alpha_{*2})$$

and $\hat{\sigma}_{22} = n^{-1}\sum_{i=1}^n \pi(\widetilde{\boldsymbol{\xi}}^\top \mathbf{z}_i)\{1-\pi(\widetilde{\boldsymbol{\xi}}^\top \mathbf{z}_i)\}\{\tilde{C}_2^\top(\mathbf{x}_i, \mathbf{y}_i)\}^2$, where

$$\tilde{C}_1(\mathbf{x}, \mathbf{y}, \alpha_*) = \{1-2\pi(\widetilde{\boldsymbol{\xi}}_*^\top \mathbf{z})\}\mathbf{y}^\top \mathbf{y} - \widetilde{\mathbf{H}}(\alpha_*)\widetilde{\mathbf{S}}_{e,x}^{-1}\mathbf{z}_e$$

and $\tilde{C}_2(\mathbf{x}, \mathbf{y}) = \mathbf{y}^\top \mathbf{1} - \widetilde{\mathbf{S}}_{xy}^\top \widetilde{\mathbf{S}}_x^{-1}\mathbf{z}$. We can straightforwardly verify that for any two constants α_{*1} and α_{*2} with consistent estimators $\hat{\alpha}_{*1}$ and $\hat{\alpha}_{*2}$, we have that $\hat{\sigma}_{11}(\hat{\alpha}_{*1}, \hat{\alpha}_{*2})$ and $\hat{\sigma}_{22}$ are consistent estimators of $\sigma_{11}(\alpha_{*1}, \alpha_{*2})$ and σ_{22} whether the data are retrospective or prospective.

4 | PROPOSED SCORE TESTS

Let the standardized score tests for random and fixed effects be $U_{1s}(\alpha_*) = n^{-1/2}U_1(\alpha_*)/\sqrt{\hat{\sigma}_{11}(\alpha_*, \alpha_*)}$ and $U_{2s} = n^{-1/2}U_2/\sqrt{\hat{\sigma}_{22}}$. Theorem 1 and the consistency of the variance estimators imply that $U_{1s}(\alpha_*)$ (for fixed α_*) and U_{2s} are asymptotically independent and have an asymptotically standard normal distribution. In this section, all limits are taken under $H_0 : \theta = 0 \;\&\; \gamma = 0$.

Since the hypothesis with respect to γ is two-sided, we reject $\gamma = 0$ if $\text{FS} = U_{2s}^2$ is large enough. The hypothesis for θ is one-sided ($\theta \ge 0$) and a larger $U_{1s}(\alpha_*)$ supports $\theta > 0$, so we reject $\theta = 0$ for large values of $\text{RS}(\alpha_*) = \{U_{1s}^+(\alpha_*)\}^2$, where $U_{1s}^+(\alpha_*) = \max\{U_{1s}(\alpha_*), 0\}$. To capture the nonnull hypothesis in both fixed and random effects, we take both scores into account and define

$$\text{SS}(\alpha_*) = \{U_{1s}^+(\alpha_*)\}^2 + U_{2s}^2. \tag{10}$$

It is worth pointing out that the tests $\text{SS}(\hat{\alpha})$ and $\text{RS}(\hat{\alpha})$ correspond to prospective score tests whereas $\text{SS}(\alpha_p)$ and $\text{RS}(\alpha_p)$ correspond to retrospective score tests. As n goes to infinity, the limiting distributions of $\text{FS}, \text{RS}(\alpha_*)$, and $\text{SS}(\alpha_*)$ are χ_1^2, $0.5\chi_0^2 + 0.5\chi_1^2$, and $0.5\chi_1^2 + 0.5\chi_2^2$, respectively. Moreover, our combination of score statistics differs from the conventional linear combination of score tests with normally distributed limiting distributions.

The parameter α_p is generally unknown in practice and the case-control data contain no information about p or α_p. If we have a guess for α_p, such as α_{*i} ($i = 1, \ldots, m$), we can define another two tests: $\text{RS}(\alpha_{*1}, \ldots, \alpha_{*m}) = \max_{1\le i\le m}\text{RS}(\alpha_{*i})$ and

$$\text{SS}(\alpha_{*1}, \ldots, \alpha_{*m}) = \max_{1\le i\le m}\text{SS}(\alpha_{*i}). \tag{11}$$

By Theorem 1, we have that $\text{RS}(\alpha_{*1}, \ldots, \alpha_{*m})$ converges in distribution to $\max_{1\le i\le m}(Z_i^+)^2$, where $(Z_1, Z_2, \ldots, Z_m)$ follows an m-variate normal distribution with mean zero and variance

$$\boldsymbol{\Sigma}_s(\alpha_{*1}, \ldots, \alpha_{*m}) = \left(\sigma_{11}(\alpha_{*i}, \alpha_{*j})/\sqrt{\sigma_{11}(\alpha_{*i}, \alpha_{*i})\sigma_{11}(\alpha_{*j}, \alpha_{*j})}\right)_{1\le i,j\le m}.$$

If we let $F((t_1,\dots,t_m),\mathbf{\Sigma})$ be the distribution function of the m-dimensional normal distribution with mean zero and variance $\mathbf{\Sigma}$, it follows that as $n\to\infty$,

$$\lim_{n\to\infty} P(\mathrm{RS}(\alpha_{*1},\dots,\alpha_{*m})\le t) = F\Big(\sqrt{t}\mathbf{1},\mathbf{\Sigma}_s(\alpha_{*1},\dots,\alpha_{*m})\Big).$$

The distribution $F((t_1,\dots,t_m),\mathbf{\Sigma})$ can be calculated by the `pmvnorm` function of the `R` package `mvtnorm`. Similarly, $\mathrm{SS}(\alpha_{*1},\dots,\alpha_{*m})$ converges in distribution to $Z_0^2+\max_{1\le i\le m}(Z_i^+)^2$, where Z_0 denotes a random variable following the standard normal distribution that is independent of Z_i $(1\le i\le k)$. Straightforward calculus gives

$$\lim_{n\to\infty} P(\mathrm{SS}(\alpha_{*1},\dots,\alpha_{*m})\le t)$$
$$= 2\int_0^{\sqrt{t}} F\Big(\sqrt{t-v^2}\mathbf{1},\mathbf{\Sigma}_s(\alpha_{*1},\dots,\alpha_{*m})\Big)\phi(v)dv.$$

In practice, we may make an interval guess about the prevalence p by experience or prior information. In cancer studies, p can be retrieved from the SEER Program (2019), an authoritative source for cancer statistics in the United States. Let $[b_1,b_2]$ be an interval guess for $\log\{p/(1-p)\}$ and let $\hat{\alpha}$ be the maximum empirical likelihood estimator of α, which together with $\widehat{\boldsymbol{\beta}}$ maximizes (9). Given $m>1$, we set $\alpha_{*i}=\hat{\alpha}-\log(n_1/n_0)+(i-1)\times(b_2-b_1)/(m-1)$ and define $\mathrm{RS}([b_1,b_2],m)=\mathrm{RS}(\alpha_{*1},\dots,\alpha_{*m})$ and

$$\mathrm{SS}([b_1,b_2],m)=\mathrm{SS}(\alpha_{*1},\dots,\alpha_{*m}), \tag{12}$$

where $\mathrm{SS}(\alpha_{*1},\dots,\alpha_{*m})$ is defined in Equation (11). In our simulation study, we set $m=4$ and $[b_1,b_2]=[-10,-0.5]$, which corresponds to the case where the disease prevalence p falls in the interval $[4.54\times10^{-5},\ 0.38]$.

5 | SIMULATION

5.1 | Simulation settings

We conduct simulations to investigate the finite-sample performance (including the type I error and power) of the proposed tests. We simulated case-control data with an equal number $n_0=n_1=2000$ of cases and controls from Model (1). The covariate is $\mathbf{x}=(x_1,x_2)^\top$, where x_1 and x_2 are independently generated from a Bernoulli distribution with success probability 0.5 and $N(1,1)$, respectively. The components of the random-effect $\mathbf{v}$ are iid $N(0,1)$ random variables. The genotype values were simulated under the Hardy-Weinberg equilibrium and linkage equilibrium. In the simulation studies below, the minor allele frequency (MAF) is considered in all possible ranges, from common to rare. The MAF refers to the frequency at which the second most common allele occurs in a given population.

TABLE 1 Type I errors (%) of the tests for Example 1 at significance levels 5%, 1%, and 0.1%

	Example 1		
Level	**5**	**1**	**0.1**
SS–MAX	5.01	1.00	0.11
SS(α_p)	4.89	1.17	0.13
SS($\hat{\alpha}$)	5.05	1.10	0.11
Burden	5.03	1.11	0.16
SKAT	5.27	1.11	0.16
SKAT-O	5.29	1.34	0.24
MiST	5.02	0.97	0.12

Our simulation settings mimic those of Sun *et al.* (2013) but with two differences. One is that they considered the continuous covariate case, but our data-generating model involves both binary and continuous covariates; this difference is not essential. The other is that their simulated subjects share the same random effect if it exists, but our random effects are iid across the individuals. Unless stated otherwise, all the results in this section are based on data generated under the independent-random-effect assumption.

5.2 | Test for overall genetic effect

The primary objective of a genetic association study is to determine the existence of an overall genetic effect, not simply a fixed or random effect. Hence, we focus on testing the overall genetic effect, for which the SS test is designed. We study the finite-sample performance of the proposed SS test by comparing the following tests: (a) the burden test (Burden for short) calculated by the `R` package `SKAT`; (b) SKAT (Wu *et al.*, 2011); (c) SKAT-O (Lee *et al.*, 2012); (d) MiST (Sun *et al.*, 2013); (e) SS(α_p), where SS($\cdot$) is defined in Equation (10); (f) SS($\hat{\alpha}$); (g) SS–MAX: SS($[-10,-0.5],4$), where SS($[b_1,b_2],m$) is defined in Equation (12). The data are generated from Example 1.

Example 1. When generating the genotype vector $\mathbf{y}$, we set the MAFs of the genotypes to $\mathrm{MAF}_j=j/(3q+1)$ $(j=1,2,\dots,q)$. We set the dimension of $\mathbf{v}$ and $\mathbf{y}$ to $q=10$, and set $\alpha_p=-1$ and $\boldsymbol{\beta}=(0.5,-1)^\top$. We consider four scenarios: (C1) $\sqrt{\theta}=0$, $\boldsymbol{\gamma}=(-0.02k)\times\mathbf{1}_q$; (C2) $\sqrt{\theta}=0.5$, $\boldsymbol{\gamma}=(-0.02k)\times\mathbf{1}_q$; (C3) $\sqrt{\theta}=0.15k$, $\boldsymbol{\gamma}=0\times\mathbf{1}$; (C4) $\sqrt{\theta}=0$, $\boldsymbol{\gamma}=(0.05k)\times\boldsymbol{\gamma}_0$, where $\boldsymbol{\gamma}_0=(\mathbf{1}_{q/2}^\top,-\mathbf{1}_{q/2}^\top)^\top$ and $k=0,1,\dots,5$. The disease prevalence is between 0.15 and 0.25.

We first compare the tests in terms of the type I errors. Table 1 gives the simulated type I errors of the seven tests (as percentages) at the nominal levels 5%, 1%, and 0.1%. The results are obtained from 10,000 replications. We can see that the proposed SS tests as well as the other four tests have well-

TABLE 2 Rejection rates (%) of the seven tests for overall genetic effect in Scenarios (C1)–(C4) of Example 1 ($\boldsymbol{\gamma}_0 = (\mathbf{1}_{q/2}^\top, -\mathbf{1}_{q/2}^\top)^\top$)

	(C1) $\sqrt{\theta} = 0, \boldsymbol{\gamma} = (-0.02k) \times \mathbf{1}_q$				
	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
SS-MAX	12.00	36.10	69.60	92.80	99.20
SS(α_p)	12.40	39.45	74.10	94.45	99.35
SS($\hat{\alpha}$)	12.55	39.35	73.70	94.35	99.35
Burden	5.85	10.75	16.55	27.30	40.65
SKAT	4.60	7.00	7.70	11.20	15.50
SKAT-O	5.90	9.65	14.25	22.95	35.45
MiST	11.40	36.45	69.30	91.75	98.85
	(C2) $\sqrt{\theta} = 0.5, \boldsymbol{\gamma} = (-0.02k) \times \mathbf{1}_q$				
SS-MAX	83.05	70.40	56.70	51.65	56.35
SS(α_p)	85.15	72.55	60.40	54.80	59.30
SS($\hat{\alpha}$)	78.80	61.30	44.35	37.30	41.10
Burden	8.65	6.60	5.30	5.55	7.70
SKAT	5.70	5.25	5.30	5.65	5.55
SKAT-O	8.15	6.25	5.50	5.65	7.35
MiST	53.05	23.75	11.95	5.60	7.95
	(C3) $\sqrt{\theta} = 0.15k, \boldsymbol{\gamma} = 0 \times \mathbf{1}_q$				
SS-MAX	8.00	35.10	85.15	99.25	100.00
SS(α_p)	8.15	38.20	86.40	99.35	100.00
SS($\hat{\alpha}$)	8.10	33.85	82.00	98.60	100.00
Burden	4.85	6.40	11.70	19.20	25.40
SKAT	4.45	4.95	6.85	7.60	10.35
SKAT-O	4.85	6.10	10.85	16.40	21.60
MiST	6.70	20.55	64.90	90.60	98.05
	(C4) $\sqrt{\theta} = 0, \boldsymbol{\gamma} = (0.05k) \times \boldsymbol{\gamma}_0$				
SS-MAX	8.45	24.75	53.20	80.85	92.90
SS(α_p)	9.45	29.60	58.15	84.60	94.80
SS($\hat{\alpha}$)	11.70	32.75	61.65	86.65	95.70
Burden	11.95	35.35	63.60	85.95	95.95
SKAT	6.60	15.35	31.70	53.50	78.35
SKAT-O	10.65	30.20	58.35	82.35	94.80
MiST	23.35	87.55	99.85	100.00	100.00

controlled type I errors even when the nominal level is just 0.1%.

We next compare the tests in terms of power. Table 2 gives the simulated powers of the seven tests at the 5% level under four scenarios. Model (1) is correct in the first three scenarios but violated in the last scenario. The results are based on 2000 replications. In Scenario (C1), SS(α_p), SS($\hat{\alpha}$), SS–MAX, and MiST have almost the same power and are clearly much more powerful than the other tests. In Scenarios (C2) and (C3), the independent random-effect model assumption is satisfied. The three SS tests outperform the other tests: in most cases, the power differences are quite significant. In Scenario (C4), Model (1) is violated and there is a fixed effect but no random effect. In this setting, MiST has the largest power, while the SS tests are comparable with Burden and SKAT-O and more powerful than SKAT.

Since SKAT, SKAT-O, and MiST are all built on the common-random-effects assumption, it is interesting to examine how the tests perform when this assumption is satisfied. We have repeated the simulations for Scenarios (C2) and (C3) of Example 1, except that the data were generated under the common-random-effects assumption. The results are reported in the Supporting Information. The SS tests are still more powerful than the Burden test. It is not surprising that they tend to be less powerful than SKAT, SKAT-O, and MiST, which are designed under this assumption.

It is worth mentioning that in Scenario (C2), SS(α_p), the SS test with the true prevalence, is much more powerful than

TABLE 3 Test results for the top six genes, in which at least one of the SS-MAX, SS($\hat{\alpha}$), Burden, SKAT, SKAT-O, and MiST tests produces a P-value less than10^{-4}; the smallest p-value of each test over the top six genes is marked in bold

Gene name	IRX2	CAMK2N1	QPCTL	BCAR1	CTNNA2	ZDHHC11B
No. of SNPs	107	32	18	20	9	13
SS-MAX	$\mathbf{7.00 \times 10^{-6}}$	3.30×10^{-3}	4.88×10^{-2}	4.94×10^{-2}	0.19	0.99
SS($\hat{\alpha}$)	1.06×10^{-4}	$\mathbf{1.85 \times 10^{-5}}$	3.55×10^{-2}	3.68×10^{-2}	0.76	0.82
Burden	7.5×10^{-2}	0.38	5.00×10^{-5}	2.20×10^{-3}	$\mathbf{8.70 \times 10^{-6}}$	0.72
SKAT	0.46	0.67	8.83×10^{-5}	1.50×10^{-4}	$\mathbf{8.70 \times 10^{-6}}$	0.73
SKAT-O	0.13	0.54	4.43×10^{-5}	2.32×10^{-4}	$\mathbf{8.70 \times 10^{-6}}$	0.74
MiST	3.64×10^{-5}	0.29	3.51×10^{-2}	$\mathbf{1.50 \times 10^{-6}}$	4.52×10^{-5}	2.49×10^{-5}

SS($\hat{\alpha}$), which is based on the score tests from the prospective likelihood. The only difference between the two tests is that they use different values of α_* in SS(α_*). Note that $\hat{\alpha} = \alpha_p + \log\{(1-p)/p\} - \log\{(1-n_1/n)/(n_1/n)\} + o_p(1)$ is nonconsistent for and different from α_p unless $n_1/n = p + o_p(1)$. The substantially different performance of SS(α_p) and SS($\hat{\alpha}$) therefore arises because the disease prevalence p (between 0.15 and 0.25) is not close to $n_1/n = 0.5$. Since α_p can be estimated from case-control data if the prevalence is known, the larger power of SS(α_p) indicates that knowledge of the disease prevalence can indeed be used to enhance the overall power.

We also conducted simulations in scenarios with low prevalence and/or rare variants, and the findings were similar. More details can be found in the Supporting Information. Overall, our score tests have desirable type I errors and are generally more powerful than existing tests when Model (1) is correct. The comparison of SS(α_p) and SS($\hat{\alpha}$) implies that prevalence information can indeed help to increase the power. The SS-MAX test is always as powerful as the ideal SS test SS(α_p), which uses the true prevalence information. We recommend SS-MAX since it may have a power gain over SS($\hat{\alpha}$) when the proportion of cases in the data is far from the prevalence. SS($\hat{\alpha}$) is also a possibility since it is usually as powerful as SS-MAX and its computational time is much lower.

6 | APPLICATION TO GWAS OF PANCREATIC CANCER

We demonstrate the performance of our method by applying it to two GWASs for pancreatic cancer. The first GWAS (PanScan I) genotyped about 550,000 SNPs from 1896 individuals with pancreatic cancer and 1939 controls drawn from 12 prospective cohorts and one hospital-based case-control study (Amundadottir *et al.*, 2009). The second GWAS (PanScan II) genotyped about 620,000 SNPs in 1679 cases and 1725 controls from seven case-control studies (Petersen *et al.*, 2010). We focused on people of predominantly European ancestry, that is, with a European admixture coefficient above 0.85 as estimated by STRUCTURE (Pritchard *et al.*, 2000). This gave 3275 cases and 3376 controls for our analysis.

We conducted gene-based multiple locus analysis on the combined data from the two GWASs. We focused on genes in the PredictDB Data Repository that were defined by eQTL SNPs identified by prediction models trained on gene expression on pancreatic tissues using GTEx Version 7 data (Gamazon *et al.*, 2015). There were 4573 genes in the data repository. We considered only the eQTL SNPs identified in the prediction model for a given gene because these SNPs have *cis* effects on the expression of the corresponding gene and thus are more likely to be functional.

We adjusted the logistic regression model for study, age, sex, and the 10 principal components (five from each GWAS) for the adjustment of population stratification. We focused on results from the following six tests: SS-MAX, SS($\hat{\alpha}$), Burden, SKAT, SKAT-O, and MiST. Table 3 gives the results for genes for which at least one of the six tests generates a P-value below 10^{-4}. We highlight P-values below 1.1×10^{-5}, which is the Bonferroni threshold for controlling the family-wise error at the level of 0.05 for a given test. Each test was able to identify one, but not necessarily the same, globally significant gene, and the P-values for a given gene can be quite different. SS-MAX detects the gene IRX2 with a P-value of 7×10^{-6}, which is much smaller than the P-value from MiST, 3.64×10^{-5}, which also provides evidence for the significance of this gene. SS($\hat{\alpha}$) identifies the gene CAMK2N1 with a P-value of 1.85×10^{-5}. MiST detects the gene BCAR1, while Burden, SKAT, and SKAT-O all detect the gene CTNNA2.

Figure 1 shows qq-plots for the tests. Assuming that most of the genes are not associated with the outcome, we would expect the qq-plot based on the P-values for the 4573 genes under each test to align well with the diagonal line. Figure 1 shows that SS-MAX and SS($\hat{\alpha}$) have the expected patterns in their qq-plots, but it is less clear for Burden, SKAT, SKAT-O, and MiST. In addition, the points at the upper right corner of each qq-plot correspond to the genes that are most likely to be significant. For example, the outlier point in this corner of the SS-MAX plot corresponds to the gene IRX2 reported in Table 3.

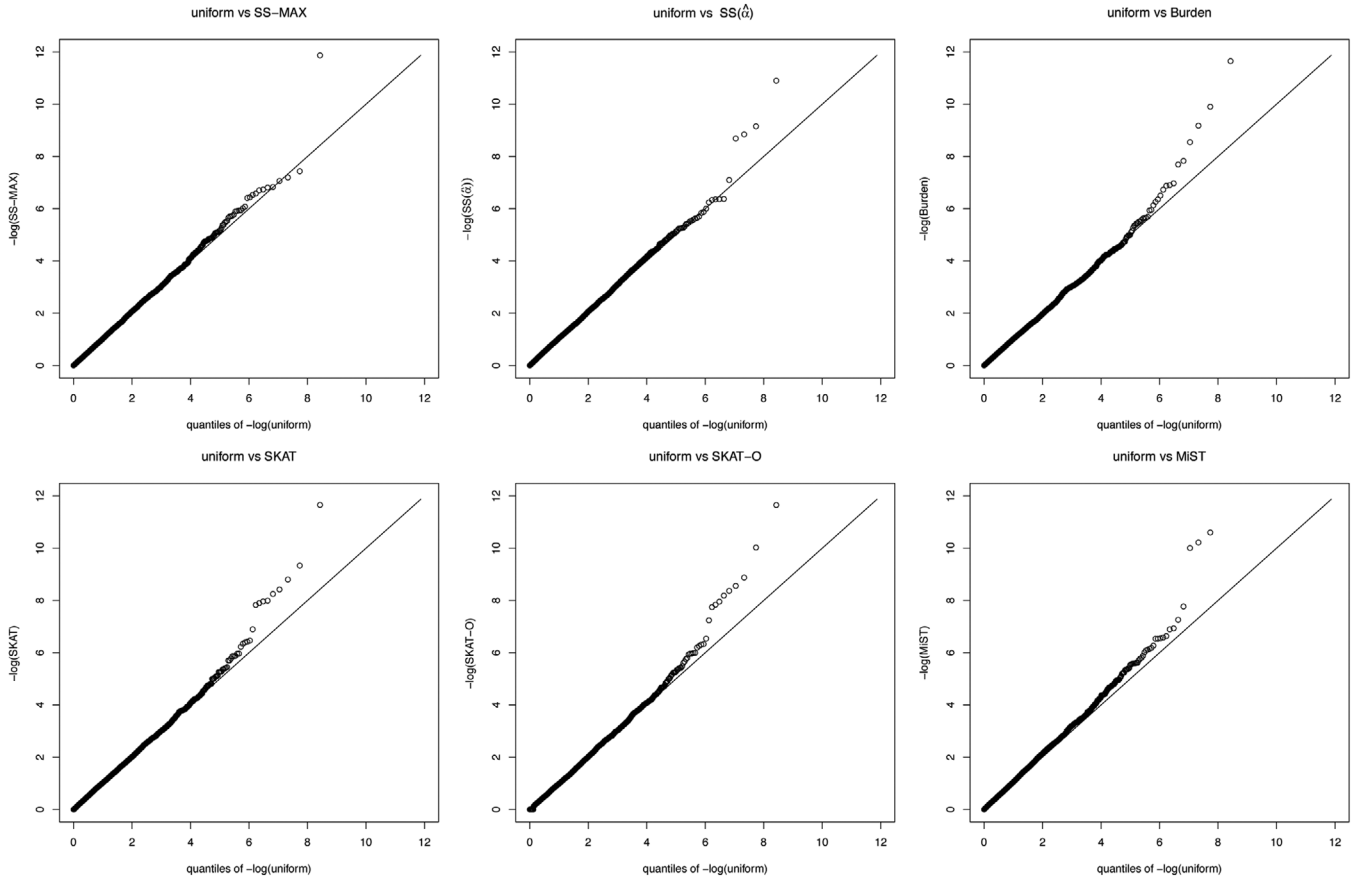

FIGURE 1 qq-plots of minus logarithm of *P*-values

7 | DISCUSSION

In this paper, we have systematically derived score-based tests for prospectively and retrospectively collected data and studied their large-sample behavior. We discussed the differences between the two likelihoods. Instead of Lin's (1997) model in which the individuals share a common random effect, we used a conventional random-effect model where the observations are independent unless they come from the same individual. Lin's (1997) common-random-effect assumption is essentially equivalent to the fixed-effect assumption, and the random effect here reflects only different data sets. We then considered data sets collected from medical centers. If we believe that the individuals in the same center share a common random effect and that different centers may have different effects generated from the same distribution, we obtain a common-random-effect assumption. On the other hand, the conventional random-effect model assumes that individuals in the same center have different effects generated from the same distribution. The conventional approach is preferable if one believes that the randomness comes from individuals rather than from centers.

In our method, the genetic information Y is linked to the disease status D only through a linear combination $\gamma \mathbf{1}^\top Y$ and only the parameter γ is unknown. The linear combination $\gamma \mathbf{1}^\top Y$ may be replaced by any combination $\gamma \mathbf{c}^\top Y$ with a user-specified direction $\mathbf{c}$. Therefore, the dimension of Y is not an issue, and our method can be directly applied to data of any dimension. To improve the power, one could consider a number of candidate directions and take the maximum of the corresponding score test as the final test statistic.

ACKNOWLEDGMENTS

The authors thank the editor, associate editor, and a referee for constructive comments and suggestions that led to significant improvements. Dr. Liu's research was supported by the National Natural Science Foundation of China (11771144, 11971300, 11871287), the State Key Program of the National Natural Science Foundation of China (71931004), Natural Science Foundation of Shanghai (17ZR1409000, 19ZR1420900), the development fund for Shanghai talents, the 111 project (B14019), and the Fundamental Research Funds for the Central Universities. Dr. Li was supported in part by the Natural Sciences and Engineering Research Council of Canada. The first two authors contributed equally to this work.

DATA AVAILABILITY STATEMENT

The data that support the findings in this paper are available from dbGaP (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000206.v5.p3) with accession id: phs000206.v5.p3. Restrictions apply to the availability of these data, which were used under license for the PanScan study.

ORCID

Yukun Liu https://orcid.org/0000-0002-9743-9276
Pengfei Li https://orcid.org/0000-0003-2165-9157
Kai Yu https://orcid.org/0000-0002-5337-137X
Jing Qin https://orcid.org/0000-0003-2817-6326

REFERENCES

Amundadottir, L., Kraft, P., Stolzenberg-Solomon, R.Z., Fuchs, C.S., Petersen, G.M., Arslan, A.A. et al. (2009) Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. *Nature Genetics*, 41, 986–990.

Anderson, J.A. (1979) Multivariate logistic compounds. *Biometrika*, 66, 17–26.

Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J. et al. (2015) A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47, 1091–1098.

Jiang, J. (2007) *Linear and Generalized Linear Mixed Models and Their Applications*. New York: Springer.

Ke, C. and Wang, Y. (2001) Semiparametric nonlinear mixed-effects models and their applications (with discussion). *Journal of the American Statistical Association*, 96, 1272–1298.

Lee, S., Wu, M. C. and Lin, X. (2012) Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13, 762–775.

Lin, X. (1997) Variance component testing in generalized linear models with random effects. *Biometrika*, 84, 309–326.

Petersen, G.M., Amundadottir, L., Fuchs, C.S., Kraft, P., Stolzenberg-Solomon, R.Z., Jacobs, K.B., et al. (2010). A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. *Nature Genetics*, 42, 224–228.

Prentice, R.L. and Pyke, R. (1979) Logistic disease incidence models and case-control studies. *Biometrika*, 66, 403–411.

Pritchard, J.K., Stephens, M. and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics*, 155, 945–959.

Qin, J. (2017) *Biased Sampling, Over-Identified Parameter Problems and Beyond*. Singapore: Springer.

Qin, J. and Zhang, B. (1997) A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika*, 84, 609–618.

SEER (2019) Surveillance, Epidemiology, and end results (SEER) program (https://www.seer.cancer.gov) research data (1975–2016). National Cancer Institute, DCCPS, Surveillance Research Program (Released April 2019, based on the November 2018 submission).

Sun, J., Zheng, Y. and Hsu, L. (2013) A unified mixed-effects model for rare-variant association in sequencing studies. *Genetic Epidemiology*, 37, 334–344.

Verbeke, G. and Lesaffre, E. (1996) A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, 91, 217–221.

Wang, Y. (1998) Mixed-effects smoothing spline analysis of variance. *Journal of the Royal Statistical Society, Series B*, 60, 159–174.

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. *American Journal of Human Genetics*, 89, 82–93.

SUPPORTING INFORMATION

Web Appendices referenced in Sections 2, 3, and 5 are available with this paper at the Biometrics website on Wiley Online Library. A zip file with R code and simulated data is available from this website as well.

Supporting Information

How to cite this article: Liu Y, Li P, Song L, Yu K, Qin J. Retrospective versus prospective score tests for genetic association with case-control data. *Biometrics*. 2021;77:102–112. https://doi.org/10.1111/biom.13270