



The sequence kernel association test for multicategorical outcomes

Zhiwen Jiang¹  | Haoyu Zhang² | Thomas U. Ahern² |
Montserrat Garcia-Closas² | Nilanjan Chatterjee³ | Hongtu Zhu¹ |
Xiang Zhan⁴ | Ni Zhao³ 

¹Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

²Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, USA

³Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland, USA

⁴Department of Biostatistics, Peking University, Beijing, China

Correspondence

Xiang Zhan, Department of Biostatistics, Peking University, Beijing 100191, China.
Email: zhanx@bjmu.edu.cn

Ni Zhao, Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205, USA.
Email: nzhao10@jhu.edu

Funding information

National Institutes of Health; National Key Research and Development Program of China

Abstract

Disease heterogeneity is ubiquitous in biomedical and clinical studies. In genetic studies, researchers are increasingly interested in understanding the distinct genetic underpinning of subtypes of diseases. However, existing set-based analysis methods for genome-wide association studies are either inadequate or inefficient to handle such multicategorical outcomes. In this paper, we proposed a novel set-based association analysis method, sequence kernel association test (SKAT)-MC, the sequence kernel association test for multicategorical outcomes (nominal or ordinal), which jointly evaluates the relationship between a set of variants (common and rare) and disease subtypes. Through comprehensive simulation studies, we showed that SKAT-MC effectively preserves the nominal type I error rate while substantially increases the statistical power compared to existing methods under various scenarios. We applied SKAT-MC to the Polish breast cancer study (PBCS), and identified gene *FGFR2* was significantly associated with estrogen receptor (ER)+ and ER- breast cancer subtypes. We also investigated educational attainment using UK Biobank data ($N = 127, 127$) with SKAT-MC, and identified 21 significant genes in the genome. Consequently, SKAT-MC is a powerful and efficient analysis tool for genetic association studies with multicategorical outcomes. A freely distributed R package SKAT-MC can be accessed at <https://github.com/Zhiwen-Owen-Jiang/SKATMC>.

KEYWORDS

multicategorical data, SKAT, the generalized logit model, the proportional odds model

1 | INTRODUCTION

The last 15 years have observed a tremendous success of genome-wide association studies (GWAS) which have collectively identified over 55,000 unique genetic loci for nearly 5000 diseases and traits (MacArthur et al., 2017).

Case-control study has been the mainstream study design of GWAS, in which study participants are classified as “diseased” and “nondiseased” groups and tested for association with genetic variants (e.g., SNPs). However, this is an oversimplification of reality. Disease phenotypes and clinical characteristics naturally have various

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2023 The Authors. *Genetic Epidemiology* published by Wiley Periodicals LLC.

properties and structures. Multicategorical outcomes (nominal and ordinal) are frequently observed when investigating complex human diseases. For example, breast cancer is commonly classified into different subtypes based on status of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor-2 (HER2) (Prat et al., 2015). Each subtype has distinct genetic risk profiles and disease prognosis (H. Zhang, Ahearn, et al., 2020). Given various subtypes of a disease, association patterns between these subtypes and genetic variants are not necessarily the same. For instance, a variant may only contribute to the risk of a specific disease subtype but not others, or the corresponding effect sizes differ from subtype to subtype. We refer this phenomenon to heterogeneity of genetic effects across subtypes, which has been observed in many recent studies (Bareche et al., 2018). Understanding the genetic underpinning of the distinct disease subtypes has stimulated great interest over recent years (Liang et al., 2015).

In this paper, we focus on set-based association analysis, in which a set of genetic variants that fall into the same gene or genomic location are combined and tested together for association. Comparing to the single variant analysis in which each variant is tested one by one followed by multiple comparison adjustment, set-based analysis aggregates information across multiple variants. It takes into account complex connections between variants, leading to discoveries with better biological interpretability, and reduces multiple comparison burden. It is particularly advantageous to analyze rare variants when the minor allele frequency (MAF) is so small that the statistical power for detecting the association for a single variant is slim (S. Lee et al., 2014).

The burden test is one of the early approaches for analyzing associations between a set of variants and a phenotype of interest (Bocher et al., 2019; B. Li & Leal, 2008; Madsen & Browning, 2009). It collapses the information across all variants in the set into a univariate burden score, and tests association between the burden score and an outcome. Because the burden score is usually a linear combination of the single score of each variant, the burden test gains power when effects of genetic variants are the same or at least in the same direction. Its performance substantially deteriorates when association directions are opposite (S. Lee et al., 2012).

In contrast, the kernel machine regression model is another big school for set-based genetic association analysis. It has the advantage to maintain a high power even when the signals of individual variants are of different directions (Davenport et al., 2018; Wu et al., 2011). In this semiparametric regression

framework, the genetic effect is modeled through a kernel similarity matrix instead of a linear combination, allowing for signals with different directions to be combined instead of being canceled out. By the connection between the kernel machine regression models and the (generalized) linear mixed models (LMM or GLMM), testing for genetic effect is equivalent to testing a variance-component in LMM or GLMM (D. Liu et al., 2007), providing a computationally efficient way for hypothesis testing across a large number of genes. Within the kernel machine regression models, the sequence kernel association test (SKAT) is one of the most widely used approaches (Wu et al., 2011). However, the original SKAT only allows for continuous and binary outcomes. When the outcome is multicategorical, binning of multiple groups is necessary, such as grouping all subtypes of a disease into a “diseased” category, which will inevitably lead to power loss if the genetic effects are different across subtypes.

Some recent studies have tried to extend SKAT to handle multiple subtypes of diseases. Davenport et al. extended SKAT for testing multivariate binary outcomes (Davenport et al., 2018), which is statistically different from a multicategorical outcome. For a multicategorical outcome, each patient belongs into a single group; for a multivariate binary outcome, each patient may fall into multiple groups. From a practice standpoint, multicategorical outcome is a more natural characterization of disease subtypes. For multicategorical outcomes, a few efforts have been made recently. H. Zhang et al. (2021) employed a mixed-effect two-stage polytomous model score test (MTOPT) to handle multiple genotypes and multiple disease characteristics simultaneously, which focuses on correlation between tumor features, missing data, and increasing degree-of-freedom in the underlying tests of associations. The essential difference lies in that it is not a set-based test. Bocher et al. (2021) extended SKAT to multicategory outcomes using a test statistic that is analogous to the model sum of squares in Fisher's one-way analysis of variance (ANOVA). An improved moment matching method is employed to calculate p -values (H. Liu et al., 2009). For small samples (less than 2000), statistics moments are calculated by permuting or bootstrapping the response residuals of the null model; and for large sample size, theoretical moments are computed. Unfortunately, this new method is very slow when sample size is greater than 2000 (data not shown), and thus it is computationally intractable for modern GWAS. Methods proposed by M. Liu et al. (2021) and He et al. (2021) utilized generalized logit models for multicategorical outcomes, which share some similarities to one of the methods proposed in this paper (SKAT-MCN, as introduced in the next paragraph). However, they

started their approaches from a multinomial logistic regression framework and applied their models to somatic mutations, completely different from our context.

In this paper, we propose SKAT-MC, a kernel machine-based score test for testing the association between a multicategorical outcome and a group of genetic variants. SKAT-MC consists of SKAT-MCN and SKAT-MCO, where “N” stands for nominal while “O” stands for ordinal. SKAT-MCN is constructed under the generalized logit mixed model and SKAT-MCO is based on the proportional odds mixed model, both of which are classical GLMMs. To the best of our knowledge, SKAT-MCO is the first approach testing association between a set of variants and an ordinal outcome. We show, via extensive simulations and a real data analysis example, the well-controlled type I error and improved power of SKAT-MC at stringent significance levels. We believe that SKAT-MC will be a critical component of the analysis toolbox for GWAS.

The rest of this paper is organized as follows. We briefly review the essential ideas of the generalized logit model and the proportional odds model and the kernel machine regression in Section 2, and emphasize how selection of the reference category influences statistical power for nominal outcomes. In Section 3, we address the details of simulation scheme regarding type I error rate and statistical power, where we compare SKAT-MC with the burden test and SKAT under four different scenarios regarding homogeneous/heterogeneous genetic effects and known/unknown best reference category. Section 4 exhibits results of simulation studies and a real data application to Polish breast cancer study and educational attainment in UK Biobank. Finally, we briefly discuss for conclusion in Section 5.

2 | METHODS

In this session, we will first review the generalized logit model, the proportional odds model and the kernel machine regression framework. Then, we will proceed to introduce our proposed SKAT-MC approaches, followed by investigating the impact of selecting reference category.

2.1 | The generalized logit model and the proportional odds model

Suppose we observe n independent subjects that each may fall into one of the J categories of outcomes. Let

$\mathbf{y}_i = (y_{i1}, \dots, y_{ij}, \dots, y_{iJ})'$ represent the vector of outcome for the i -th subject ($i = 1, \dots, n$). $y_{ij} = 1$ indicates that the i -th subject belongs to the j -th category and $y_{ij} = 0$ otherwise. The phenotype can be any categorical clinical outcomes (nominal or ordinal), such as subtypes of a disease and increasing levels of pain. Each subject can only belong to one category such that $\sum_{j=1}^J y_{ij} = 1$ for all i . Let $\pi_j(\mathbf{x}_i) = \Pr(y_{ij} = 1 | \mathbf{x}_i)$ be the conditional probability that subject i is of category j with $\sum_j \pi_j(\mathbf{x}_i) = 1$, where $\mathbf{x}_i = (x_{i1}, \dots, x_{im})'$ denotes the set of covariates that we want to associate \mathbf{y}_i with. If \mathbf{y}_i is nominal, without loss of generalization, we set the last category J as the reference and form the generalized logit model

$$\log \frac{\pi_j(\mathbf{x}_i)}{\pi_J(\mathbf{x}_i)} = \alpha_j + \boldsymbol{\beta}'_j \mathbf{x}_i, \quad (1)$$

for $j = 1, \dots, J - 1$. Each coordinate vector $\boldsymbol{\beta}_j = (\beta_{j1}, \dots, \beta_{jm})'$ represents the increase in log-odds of being in category j versus the reference category J resulting from a one-unit increase in the corresponding covariate, controlling for other covariates. Here, α_j and $\boldsymbol{\beta}_j$, the regression coefficients are not required to be the same for any two categories. If the categories are ordinal, the order information can be incorporated into the proportional odds model as

$$\text{logit}(\nu_j(\mathbf{x}_i)) = \log \frac{\nu_j(\mathbf{x}_i)}{1 - \nu_j(\mathbf{x}_i)} = \alpha_j + \boldsymbol{\beta}' \mathbf{x}_i, \quad (2)$$

where $j = 1, \dots, J - 1$ stand for the ordered categories and

$$\nu_j(\mathbf{x}_i) = \sum_{l=1}^j \Pr(y_{il} = 1 | \mathbf{x}_i) = \pi_1(\mathbf{x}_i) + \dots + \pi_j(\mathbf{x}_i). \quad (3)$$

Because we model the probabilities up to the j -th ordered category, ν_j is called the cumulative probability. The corresponding category (or response), defined by $\tilde{y}_{ij} = \sum_{l=1}^j y_{il}$, is called the cumulative category. The model leverages the order information by incorporating categories into the cumulative category in sequence. Each coordinate of $\boldsymbol{\beta}$ instead indicates the increment of log-odds ratio of falling into the first j categories, while holding other covariates fixed. Unlike the flexibility regarding α_j and $\boldsymbol{\beta}_j$ in the generalized logit model, $\boldsymbol{\beta}$ keeps constant across $J - 1$ logits and α_j have to be monotonically increasing in the proportional odds model.

2.2 | A kernel-based generalized logit model and proportional odds model

Let $\mathbf{G}_i = (G_{i1}, \dots, G_{ip})'$ denote the genotypes for p variants (i.e., SNPs) within a specific gene or genomic region. Usually, $G_{ik} = 0, 1$ or 2 represents the number of minor alleles at variant k for individual i . The variants can be either common or rare variants, defined by MAF using a certain threshold (e.g., 0.05 or 0.01 depending on the sample size). We relate the outcome to the covariate \mathbf{x}_i and genotype \mathbf{G}_i with the following model

$$\eta_{ij} = \alpha_j + \mathbf{x}_i' \boldsymbol{\beta}_j + h_j(\mathbf{G}_i), \quad (4)$$

where $i = 1, \dots, n, j = 1, \dots, J - 1, \eta = g(\cdot)$, and $g(\cdot)$ is a link function in GLMMs. For the generalized logit model, $g(\pi_{ij}) = \log(\pi_{ij}/\pi_{iJ})$. We here arbitrarily choose the J -th category as the reference to develop association analysis and the topic of selecting reference will be discussed later. For the proportional odds model $g(\nu_{ij}) = \log\{\nu_{ij}/(1 - \nu_{ij})\}$ and all the vectors of $\boldsymbol{\beta}_j$ and $h_j(\cdot)$ are the same due to the proportional odds assumption. In both models, $h_j(\cdot)$ are unknown real functions corresponding to the genetic effects on the j -th category (nominal outcomes) or j -th cumulative category (ordinal outcomes).

In this paper, we aim at testing whether the set of variants has effect on any nonbaseline category compared to the reference category (if the outcome is nominal), or whether the effect is different for higher-ranked categories compared to the lower-ranked categories (for ordinal outcomes). The effect is fully characterized by the function $h_j(\cdot)$. Therefore, the null hypothesis corresponds to $H_0 : h_1(\cdot) = h_2(\cdot) = \dots = h_{J-1}(\cdot) = 0$. Here, like in many other kernel-based approaches, we assume that each $h_j(\cdot)$ lies in a reproducing kernel Hilbert spaces (RKHS) spanned by the positive-definite kernel function $K_j(\cdot, \cdot)$. Per Mercer's theorem (Cristianini & Shawe-Taylor, 2000), under some regularity conditions, the kernel function $K_j(\cdot, \cdot)$ specifies a unique Hilbert space and fully determines the function $h_j(\cdot)$. Moreover, any function $h_j(\mathbf{G}_i)$ can be expressed by $\sum_{i'=1}^n K_j(\mathbf{G}_i, \mathbf{G}_{i'}) a_{ji'}$, for some constants a_{j1}, \dots, a_{jn} . Using this dual representation, it is convenient to construct $h_j(\cdot)$ by the kernel function $K_j(\cdot, \cdot)$ without the need to specify its functional form. In matrix format, we specify a kernel matrix \mathbf{K}_j which measures the pairwise similarity in the genetic data with its i, i' -th element being $K_j(\mathbf{G}_i, \mathbf{G}_{i'})$. We employ the weighted/unweighted linear kernel, the most commonly used kernels in genetic association analysis (Bocher et al., 2021; Wu et al., 2011). For common variants, we use the linear kernel $K_j(\mathbf{G}_i, \mathbf{G}_{i'}) = \sum_{k=1}^p G_{ik} G_{i'k}$ corresponding to

$h_j(\mathbf{G}_i) = \sum_{k=1}^p G_{ik} b_{jk}$, where $b_{jk} = \sum_{i'=1}^n G_{i'k} a_{ji'}$. For rare variants, we use the weighted linear kernel $K_j(\mathbf{G}_i, \mathbf{G}_{i'}) = \sum_{k=1}^p w_{jk} G_{ik} G_{i'k}$. Following Wu et al. (2011), we set weights $\sqrt{w_{jk}} = \text{dbeta}(\text{MAF}_k; 1, 10)$ or $\sqrt{w_{jk}} = \text{dbeta}(\text{MAF}_k; 1, 25)$ depending on the balance of samples and the number of categories (more on this later), where $\text{dbeta}(\text{MAF}_k; \cdot, \cdot)$ is a beta distribution density function evaluated at MAF_k . This approach gives a higher weight to a rarer variant.

2.3 | A kernel association score test for multicategorical outcomes

In matrix language, the model in Equation (4) can be rewritten as

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{h}(\mathbf{G}), \quad (5)$$

where $\boldsymbol{\eta} = (\eta_{11}, \eta_{21}, \dots, \eta_{n1}, \eta_{12}, \dots, \eta_{n2}, \dots, \eta_{1, J-1}, \dots, \eta_{n, J-1})'$, a stacked column vector of the so-called "linear" predictor. $\boldsymbol{\beta} = (\alpha_1, \boldsymbol{\beta}'_1, \alpha_2, \boldsymbol{\beta}'_2, \dots, \alpha_{J-1}, \boldsymbol{\beta}'_{J-1})'$ is a vector of the regression coefficients for other covariates. $\mathbf{h}(\mathbf{G}) = (h_1(\mathbf{G}_1), \dots, h_1(\mathbf{G}_n), \dots, h_{J-1}(\mathbf{G}_1), \dots, h_{J-1}(\mathbf{G}_n))'$ represents

the genetic effect. $\mathbf{X} = \mathbf{I}_{J-1} \otimes \begin{bmatrix} 1 & \mathbf{x}'_1 \\ 1 & \mathbf{x}'_2 \\ \vdots & \vdots \\ 1 & \mathbf{x}'_n \end{bmatrix}$, with \otimes representing

Kronecker product. The null hypothesis in turn transforms to $H_0 : \mathbf{h}(\mathbf{G}) = \mathbf{0}$. Through the relationship between the kernel machine regression and mixed models (D. Liu et al., 2007), we can consider $\mathbf{h}(\mathbf{G})$ as a random effect following $\mathcal{N}(\mathbf{0}, \boldsymbol{\tau}\mathbf{K})$, where $\mathbf{K} = \text{diag}\{\mathbf{K}_1, \dots, \mathbf{K}_{J-1}\}$ with each element \mathbf{K}_j being an $n \times n$ kernel matrix. In practice, we assume that different categories share the common variance-covariance matrix, that is, $\mathbf{K}_1 = \mathbf{K}_2 = \dots = \mathbf{K}_{J-1}$, but a different kernel for each category is also theoretically allowed.

Under the mixed model framework, we derive the variance-component score test (Supporting Information: Appendix A). Specifically, the test statistic has the general form

$$Q = (\mathbf{y}^* - \mathbf{X}\hat{\boldsymbol{\beta}})' \mathbf{W}\mathbf{K}\mathbf{W} (\mathbf{y}^* - \mathbf{X}\hat{\boldsymbol{\beta}}), \quad (6)$$

where \mathbf{y}^* is the working response vector of length $n(J - 1)$ and is organized in the same way as $\boldsymbol{\eta}$ in Equation (5). \mathbf{W} is an $n(J - 1) \times n(J - 1)$ matrix denoting the working weight in GLM framework, which is a block matrix with dimension $J - 1$, with each entry being an $n \times n$ diagonal matrix. Let Q_N and Q_O represent the test statistics for SKAT-MCN (nominal data under the generalized logit model) and SKAT-MCO (ordinal

data under the proportional odds model), respectively. The score test statistics within the two models reduce to:

$$Q_N = (\mathbf{y} - \hat{\boldsymbol{\pi}})'(\mathbf{D}\mathbf{V})^{-1}\mathbf{K}(\mathbf{V}\mathbf{D})^{-1}(\mathbf{y} - \hat{\boldsymbol{\pi}})$$

and $Q_O = (\tilde{\mathbf{y}} - \hat{\boldsymbol{\nu}})' \mathbf{K}(\tilde{\mathbf{y}} - \hat{\boldsymbol{\nu}}),$

respectively, where \mathbf{y} is the original observed response and $\tilde{\mathbf{y}}$ is the cumulative response with same form as \mathbf{y} . $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\nu}}$ are the fitted values of \mathbf{y} and $\tilde{\mathbf{y}}$ from the generalized logit model and the proportional odds model under the null hypothesis, respectively. For SKAT-MCN, $\mathbf{W} = \mathbf{D}\mathbf{V}\mathbf{D}$, $\mathbf{D} = \partial\boldsymbol{\eta}/\partial\boldsymbol{\pi}$ and \mathbf{V} is the variance-covariance matrix of the multinomial distribution (assuming the dispersion parameter $\phi = 1$). Here, $\mathbf{D} = \text{diag}(\mathbf{D}_1, \dots, \mathbf{D}_{J-1})$, where $\mathbf{D}_j = \partial\boldsymbol{\eta}_j/\partial\boldsymbol{\pi}_j$ are $n \times n$ diagonal matrices. \mathbf{V} has the same form as \mathbf{W} , where each entry of the block matrix is the (co-)variance between category j and l , $j, l = 1, \dots, J-1$. That is, $\mathbf{V}_{jj} = \text{Cov}(\mathbf{y}_j, \mathbf{y}_j) = \text{diag}(\pi_{1j}(1 - \pi_{1j}), \dots, \pi_{nj}(1 - \pi_{nj}))$ and $\mathbf{V}_{jl} = \text{Cov}(\mathbf{y}_j, \mathbf{y}_l) = \text{diag}(-\pi_{1j}\pi_{1l}, \dots, -\pi_{nj}\pi_{nl}), j \neq l$. To calculate test statistics, we can substitute \mathbf{D} and \mathbf{V} with their consistent estimates obtained under the null model. Further details about these statistics along with p -value calculation are provided in Supporting Information: Appendix A, available at *Genetic Epidemiology* online.

2.4 | Selection of the reference category for SKAT-MCN

Analysis of multicategorical data often proceeds by assigning a category as the reference and then comparing the rest of categories with it. For ordinal data, the order of categories dictates that we may only choose the first or the last category as the reference. Because the proportional odds model dichotomizes the outcome categories using cumulative outcomes (i.e., categories 1 to j vs. categories $j+1$ and above), accumulating ordered categories forward or backward would produce the same result. However, for nominal data, we enjoy the freedom of choosing any category as the reference, which gives rise to a natural question of how to choose an appropriate reference in the SKAT-MCN framework. Indeed, for SKAT-MCN, choosing different reference categories will generate distinct results. Overall, we observed that the type I error is always well-controlled no matter which reference category is chosen since all tests share the same null hypothesis of no differences among all categories. However, the statistical power of tests can differ substantially with different references being selected since the alternative hypotheses are different when

shifting the reference category. These phenomena can be explained by carefully investigating the hypotheses under changes of references.

Taking Equation (4) as an example, if we set the last category as the reference, $\mathbf{h}_j = \mathbf{0}$ entails that there is no difference in the genetic data between categories j and the reference category J . When we change the reference, suppose to the category 1, the model becomes $\log \frac{\pi_j(x_i)}{\pi_1(x_i)} = \alpha_j^* + \boldsymbol{\beta}_j^{*'} \mathbf{x}_i + \mathbf{h}_j^*$, $i = 1, \dots, n, j = 2, \dots, J$. Note that \mathbf{h}_j^* has a different interpretation: $\mathbf{h}_j^* = \mathbf{0}$ means that there is no difference in the genetic effect between categories j and the reference category 1. Under the (complete) null $H_0: \mathbf{h}_1 = \dots = \mathbf{h}_{J-1} = \mathbf{0}$, there is no association at all between the genetic variants and any category, which is equivalent to $H_0^*: \mathbf{h}_2^* = \dots = \mathbf{h}_J^* = \mathbf{0}$. Therefore, the type I error should be preserved no matter which reference is chosen. However, this equivalence no longer holds under the alternative. Because we compare all categories with the reference, consider a scenario that there is no genetic difference between categories 1 to $J-1$, but category J is genetically very different from the others. Then if we choose category J as the reference, every one of $\mathbf{h}_1, \dots, \mathbf{h}_{J-1}$ is of the same nonzero value. However, if we choose category 1 as the reference, $\mathbf{h}_2^*, \dots, \mathbf{h}_{J-1}^*$ are all zeros, with only $\mathbf{h}_J^* \neq \mathbf{0}$. In this situation, choosing category J as the reference will tend to be more powerful. Generally speaking, setting the category that has the largest genetic disparity (for the variant-set) as the reference will achieve the highest power.

A recent study on this problem provides an alternative explanation that category with the weakest correlation with variants should be treated as the reference (He et al., 2021), as it is often the case that corresponds to the largest test statistic. It is widely known that the statistical power of a test is determined by both the alternative hypothesis and test statistic. Changing the reference category would have an impact on both the alternative hypothesis and test statistic, and it is challenging to mathematically figure out which factor has a larger influence on the statistical power. In real studies, it is typically unknown which category satisfies the aforementioned criteria, and even for the same study, the best reference may differ from gene to gene because of the distinct genetic effects. In this paper, in the lack of external information, we propose to treat each category as the reference one by one and then use Cauchy combination (Y. Liu & Xie, 2020) to aggregate the p -values from individual tests using each category as reference. This approach eliminates the need to choose a reference and is statistically robust: it suffers from slight power loss compared to the best scenario, but will gain substantial power compared with a poor choice of reference.

3 | SIMULATION STUDIES

We conducted comprehensive simulations to evaluate the performance of SKAT-MC with respect to type I error rate and statistical power, for common and rare variants separately. We utilized hapgen2 (Su et al., 2011) to simulate genetic data of common variants, where we used the entire chromosome 1 of Hapmap 3 (release 2) haplotypes as the reference data. The overall simulated genetic data contained 99,535 common variants (MAF > 0.05) with a total population size of 10,000. We used simuG (Yue & Liti, 2019) to generate rare variants, where the template was a randomly selected 1 mb genome section of chromosome 1 from 10,000 subjects in UK Biobank (<http://www.ukbiobank.ac.uk/resources/>). The simulated genetic data had 3,201 rare variants ($0.001 < \text{MAF} < 0.05$) and 10,000 subjects. When simulating genetic data for a particular sample size, we randomly chose a 30 kb (resp. 10 kb) section for common variants (resp. rare variants).

3.1 | Type I error simulations

To evaluate the empirical type I error rate of SKAT-MC at a genome-wide significance level (e.g., $\alpha = 2.5 \times 10^{-6}$ as widely used for gene-based association analysis), we generated 10^8 replicates under the null model. However, due to the computational burden, we actually generated 10,000 genotype matrices and 10,000 null models. Then the empirical type I error rate was computed from the p -values of 10^8 genotype-and-null model combinations. Specifically, we generated covariates \mathbf{X} and response vectors \mathbf{y} through the null model

$$g(E(y_{ij})) = \alpha_j + 0.5x_{i1} + 0.5x_{i2}, \quad (7)$$

where $i = 1, 2, \dots, n$ and $n = 1000, 2500, 5000$ for various sample sizes; $j = 1, 2$ or $j = 1, 2, 3, 4$ for scenarios of three or five categories, respectively. x_{i1} is a continuous variable generated from a standard normal distribution and x_{i2} is an indicator variable generated from a Bernoulli distribution with probability 0.5. For nominal data, $g(\cdot)$ is the link function of the generalized logit model and $\alpha_j = -4$ for common variants and $\alpha_j = -1$ for rare variants. Under the null with no genetic effects on groups, each reference group is equivalent for SKAT-MCN. For ordinal data, $g(\cdot)$ is the link function of the proportional odds model and $\alpha_j = j - 5$ for common variants and $\alpha_j = j - 2$ for rare variants. The outcome y_{ij} were simulated from the categorical distribution with the probability calculated from the inverse link functions of the generalized logit model and the proportional odds model, respectively. The

unweighted linear kernel was applied to common variants. In contrast, the weighted linear kernel with weight $\text{dbeta}(\text{MAF}, 1, 25)$ was applied to rare variants when there are three categories and $\text{dbeta}(\text{MAF}, 1, 10)$ when there were five categories. The different weights used for different number of categories were dedicated since we found that the weight could affect type I error (more on this in the result and discussion section). The empirical type I error rates were determined by the proportion of p -values less than several significance levels ($\alpha = 0.05, 10^{-4}, 2.5 \times 10^{-6}$, and 10^{-6}).

3.2 | Statistical power simulations

We considered two important factors that reflect the nature of multicategorical data: (1) knowing the true reference category or not; and (2) homogeneous/heterogeneous genetic effects across categories (e.g., the genetic variants may affect only a specific subtype of disease or even have effects with different directions and sizes on different subtypes). As a result, we generated four simulation scenarios (Table 1). These scenarios were applied to both common and rare variants. For Scenario 1, we simulated nominal and ordinal data, with the true reference being known for nominal data, and the genetic effects were homogeneous across subtypes (i.e., a variant has identical effect on all subtypes). For Scenario 2, we simulated nominal and ordinal data, with the true reference being known for nominal data, and the genetic effects were heterogeneous (i.e., a variant may have different effects on different subtypes). For Scenario 3, we simulated nominal data only with unknown true reference and homogeneous effects. For Scenario 4, we simulated nominal data only with unknown true reference and heterogeneous effects.

Considering the case-control design in which case group may be classified into multiple subtypes, for nominal data, we simulated the same number of cases as that of controls, where cases were evenly split into multiple subtypes. For the ordinal data, we simulated equal sample size in each category. For Scenario 1, we simulated covariates and outcomes with the model

$$g(E(y_{ij})) = \alpha_j + 0.5x_{i1} + 0.5x_{i2} + b_1 G_{i1} + b_2 G_{i2} + \dots + b_p G_{ip}, \quad (8)$$

where the definition of $g(\cdot)$, x_{i1} , x_{i2} , α_j , and n are the same as before in the type I error simulation study. Out of all the simulated variants from a 30 kb section for common variants (resp. 10 kb section for rare variants), we randomly selected 30% (resp. 10%) to be causal, and

TABLE 1 Scenarios of power simulation studies.

Scenario	True reference for nominal data	Genetic effects	Data type
1	Known	Homogeneous across categories	Nominal and ordinal
2	Known	Heterogeneous across categories	Nominal and ordinal
3	Unknown	Homogeneous across categories	Nominal
4	Unknown	Heterogeneous across categories	Nominal

designated them as G_{is} , $s = 1, \dots, p_c$ as in Equation (8). The coefficient $|b_s| = 0.402|\log_{10} \text{MAF}_s|$ such that rarer SNPs have larger (absolute) effect sizes. As we can see, in this simulation setting, we implicitly used the last category as the reference to generate data as we incorporated the first $J - 1$ categories in Equation (8). Therefore, we set the last group as the reference when implementing SKAT-MCN to reflect the knowledge of best reference. Following Wu et al. (2011), we considered coefficients of causal variants b_1, \dots, b_{p_c} with 100% positive, 80% positive and 50% positive. It means variants within a genetic set have effects in different directions, but the effects are homogeneous across subtypes.

For Scenario 2, we simulated data so that genetic effects were different for different categories (category 1 vs. Category 2 when there are three categories, and Categories 1 and 2 vs. Categories 3 and 4 when there are five categories). Specifically, for the simulation with three categories, we set the last category as the reference and generate the probability for Category 1 as

$$g(E(y_{i1})) = \alpha_1 + 0.5x_{i1} + 0.5x_{i2} + b_1G_{i1} + b_2G_{i2} + \dots + b_{p_c}G_{ip_c}, \quad (9)$$

with $|b_s| = 0.402|\log_{10} \text{MAF}_s|$. Then we randomly chose half causal variants to have positive effect size $|b_s|$ and half causal variants to have negative effect size. For Category 2, we have

$$g(E(y_{i2})) = \alpha_2 + 0.5x_{i1} + 0.5x_{i2} + b_1^*G_{i1} + b_2^*G_{i2} + \dots + b_{p_c}^*G_{ip_c}, \quad (10)$$

with all $b_s^* = |b_s|$ being positive. In this way, the genetic effects were heterogeneous across the two categories. Similarly, for simulating five categories, the regression coefficients were the same for Categories 1 and 2 and for Categories 3 and 4, respectively, with half of them being negative for Categories 1 and 2, and all positives for Categories 3 and 4.

Scenario 3 was similar to Scenario 1, and Scenario 4 was close to Scenario 2. But instead of keeping the last

category as the reference as the previous scenarios did, we randomly chose 30% variants with the first category being the true reference, while for the rest, the true reference was still set as the last category. In this case, we mimicked the situation that no prior knowledge about the true reference was available. These two scenarios were only available for nominal data (SKAT-MCN).

We applied SKAT-MCN to nominal data in all four scenarios and SKAT-MCO only to ordinal data. SKAT-MCN was also applied to ordinal data because ordinal data satisfies assumptions of the generalized logit model. By doing so, we hope to illustrate that SKAT-MCO indeed is more powerful than SKAT-MCN by taking advantage of the order information. For simulation Scenarios 3 and 4, we attempted individual SKAT-MCN tests with each category being the reference and then used Cauchy combination (Y. Liu & Xie, 2020) to generate an overall omnibus p -value. As for SKAT, because it is not adequate to multicategorical data, we combined the first $J - 1$ categories together to form a “case” group and compare it with controls (the J -th category) in all scenarios. The unweighted linear kernel was used in both SKAT-MC and SKAT when dealing with common variants, and the weighed linear kernel was utilized for rare variants with weight $\sqrt{w} = \text{dbeta}(\text{MAF}, 1, 25)$ (three categories) and $\sqrt{w} = \text{dbeta}(\text{MAF}, 1, 10)$ (five categories). For the burden test, we first generated a burden score by summing up the number of minor alleles. The rule for using weights was the same as SKAT-MC. Then we fitted a generalized logit model or a proportional odds model between the burden score (fixed effect) and the outcome, and tested for association using the Wald test. We evaluated statistical power using 2000 simulations at the level $\alpha = 2.5 \times 10^{-6}$.

4 | RESULTS

4.1 | Simulation studies of type I error and statistical power

The empirical type I error rates of both SKAT-MCN and SKAT-MCO were successfully protected at different nominal significance levels $\alpha = 0.05, 10^{-4}, 2.5 \times 10^{-6}, 10^{-6}$, for

TABLE 2 Empirical type I error rates of sequence kernel association test (SKAT)-MC with three-category outcomes and common variants.

n	$\alpha = 0.05$	$\alpha = 10^{-4}$	$\alpha = 2.5 \times 10^{-6}$	$\alpha = 10^{-6}$
SKAT-MCN				
1000	5.03×10^{-2}	9.78×10^{-5}	2.50×10^{-6}	9.99×10^{-7}
2500	5.01×10^{-2}	9.78×10^{-5}	2.03×10^{-6}	7.41×10^{-7}
5000	5.00×10^{-2}	9.92×10^{-5}	2.35×10^{-6}	9.09×10^{-7}
SKAT-MCO				
1000	5.05×10^{-2}	9.71×10^{-5}	2.08×10^{-6}	8.39×10^{-7}
2500	5.03×10^{-2}	9.80×10^{-5}	2.07×10^{-6}	8.64×10^{-7}
5000	5.02×10^{-2}	9.91×10^{-5}	2.31×10^{-6}	8.41×10^{-7}

TABLE 3 Empirical type I error rates of sequence kernel association test (SKAT)-MC with three-category outcomes and rare variants.

n	$\alpha = 0.05$	$\alpha = 10^{-4}$	$\alpha = 2.5 \times 10^{-6}$	$\alpha = 10^{-6}$
SKAT-MCN				
1000	4.85×10^{-2}	7.26×10^{-5}	1.28×10^{-6}	4.20×10^{-7}
2500	4.94×10^{-2}	8.72×10^{-5}	1.66×10^{-6}	6.64×10^{-7}
5000	4.98×10^{-2}	9.33×10^{-5}	2.04×10^{-6}	7.83×10^{-7}
SKAT-MCO				
1000	4.84×10^{-2}	9.11×10^{-5}	1.24×10^{-6}	6.40×10^{-7}
2500	4.95×10^{-2}	9.23×10^{-5}	1.86×10^{-6}	7.61×10^{-7}
5000	4.98×10^{-2}	9.34×10^{-5}	2.07×10^{-6}	8.93×10^{-7}

common and rare variants and for the total sample size 1000, 2500, and 5000 with $J = 3$ (Table 2, common variants; Table 3, rare variants) and $J = 5$ (Supporting Information: Table S1 available at *Genetic Epidemiology* online) categories. We found that balance of samples across categories and weights used in the weighted linear kernels could affect the type I error. Specifically, the more severe unbalance of samples, and the higher weights were assigned to rarer variants, the higher type I error would be observed at some extreme tail, such as 1.0×10^{-6} . It also happened when all the variants in the set were extremely rare (e.g., $MAF < 0.001$), or all variants were in high linkage disequilibrium (LD) with each other. This situation is mainly because at first, when samples are severely unbalanced across categories, some categories may have only few samples, then the null distribution of test statistics based on large sample theory may be invalid. Secondly, rarer variants may dominate the test if assigning extremely high weights to them. For example, $\sqrt{w} = \text{dbeta}(MAF, 1, 25)$ assigns weight 24.41 to variant A with $MAF = 0.001$ but

assigns weight 7.30 to a variant B with $MAF = 0.05$. After constructing the kernel, the importance of variant A is boosted to 11 times to that of variant B, so that variant A will drive the association test and the null distribution of test statistics may also be changed. This situation was not specific to SKAT-MC, but could be observed for the burden test (data not shown). Based on our simulation studies, we gave a rule of thumb of selecting weights for rare variants. For three-category data, if the smallest category has more than 10% samples and all variants have $MAF > 0.001$, then $\sqrt{w} = \text{dbeta}(MAF, 1, 25)$ can be used. For data with more categories, more unbalanced samples, and/or rarer variants, $\sqrt{w} = \text{dbeta}(MAF, 1, 10)$ may be an advisable choice.

Figure 1 and Supporting Information: Figure S1 show the performance of SKAT-MC, SKAT, and the burden test on common (panel a and b) and rare (panel c and d) variants with three and five-category data, respectively, under scenario 1. As expected, in each panel, when the proportion of causal variants that were negatively associated with increased (from 0%, 20% to 50%), the burden test suffered from substantial power loss because it implicitly assumes that all variants influence the phenotype in the same direction. It had slim power for rare variants, even if the sample size was up to 5000. On the other hand, SKAT-MC and SKAT were robust to diverse effect directions and the reduction of power was small, because they leveraged the kernel matrix to capture the inconsistent genetic effects. For nominal outcomes, SKAT had slight power gain compared to SKAT-MCN: after all, the genetic effects were indeed the same across all categories and combining categories led to more efficient testing. For ordinal outcomes, SKAT-MCO surpassed all other methods by taking advantage of the order information. The power gain was more apparent for rare variants (panel d) and for five-category data (Supporting Information: Figure S1). The latter case is probably because, as the number of categories increases, the order information is increasingly important. And the efficiency of the proportional odds model compared with the generalized logit model is more apparent.

Figure 2 shows the results for simulation Scenario 2, in which the genetic effects were heterogeneous across categories and the true reference category was known (for nominal data). Compared to competitor methods, SKAT-MC showed substantial power gain for both nominal and ordinal data, for both common and rare variants and with both three and five categories. For nominal outcomes, SKAT-MCN was the most powerful; SKAT was the least powerful because it combined multiple categories that had distinct genetic effects into a single category for analysis. The burden test was more powerful than SKAT for nominal data and common variants, but

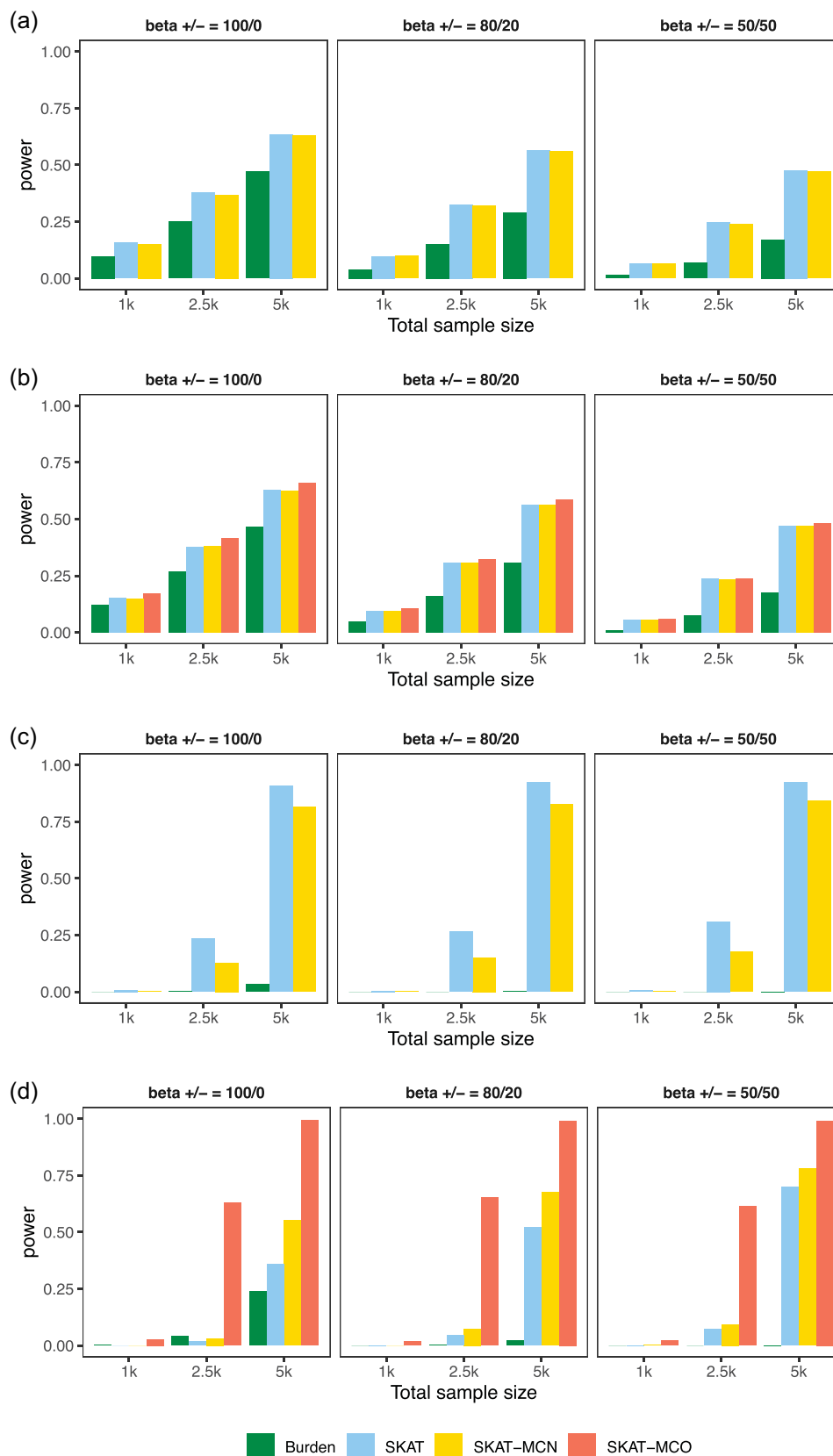


FIGURE 1 Comparisons of statistical power under simulation scenario 1: homogeneous genetic effects across categories, known reference category (for nominal outcomes) and three-category outcomes. (a) Nominal outcomes with common variants. (b) Ordinal outcomes with common variants. (c) Nominal outcomes with rare variants. (d) Ordinal outcomes with rare variants. In each panel, from left to right, the coefficients of causal variants in each gene were 100% positive (0% negative), 80% positive (20% negative), and 50% positive (50% negative), respectively. The unweighted linear kernel was applied to both sequence kernel association test (SKAT)-MC and SKAT for the common variants setting, and the weighted linear kernel with weights $\text{dbeta}(\text{MAF}, 1, 25)$ was applied to SKAT-MC, SKAT, and the burden test for the rare variants setting.

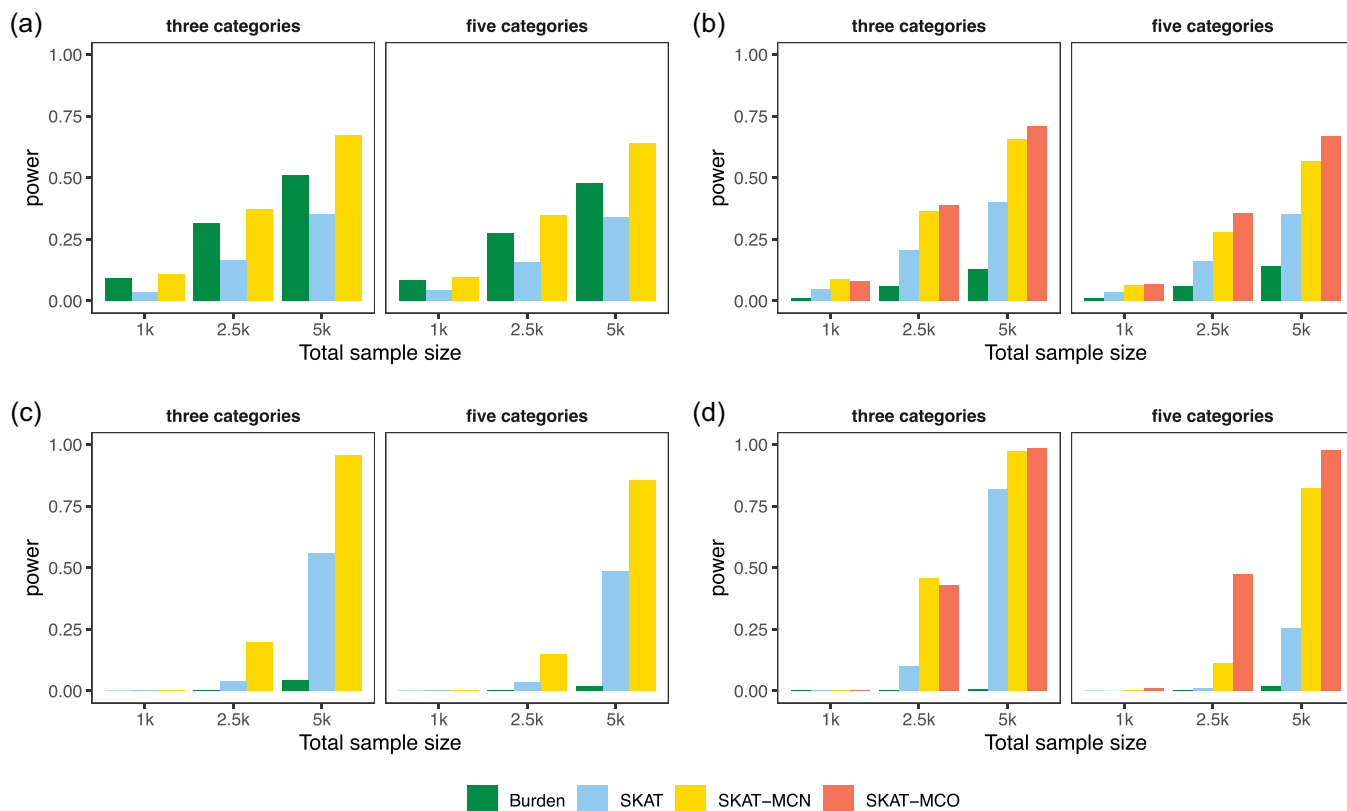


FIGURE 2 Comparisons of statistical power under simulation scenario 2: heterogeneous genetic effects across categories, known reference category (for nominal outcomes), three and five-category outcomes. (a) Nominal outcomes with common variants. (b) Ordinal outcomes with common variants. (c) Nominal outcomes with rare variants. (d) Ordinal outcomes with rare variants. The unweighted linear kernel was applied to both sequence kernel association test (SKAT)-MC and SKAT for the common variants setting, and the weighted linear kernel was applied to SKAT-MC, SKAT, and the burden test for the rare variants setting. The weights were $\text{dbeta}(\text{MAF}, 1, 25)$ for three-category data and $\text{dbeta}(\text{MAF}, 1, 10)$ for five-category data.

less powerful than SKAT-MCN (panel a). Keeping the sample size of each category the same, the statistical power are jointly determined by two factors: the degree of heterogeneous effects across categories and the proportion of variants oppositely associated with outcomes. SKAT-MCN adapted to both two factors and thus outperformed the other two approaches. For ordinal outcomes, SKAT-MCO was the most powerful, followed by SKAT-MCN, and both were much more powerful than SKAT and the burden test. The superiority of SKAT-MCO is because it was most adapted to the data. SKAT-MCN was robust to the heterogeneity of effects across categories, although it neglected the order information. SKAT and the burden test suffered, mainly because SKAT is sensitive to the heterogeneity of effects across categories and the burden test is vulnerable to high proportion of opposite effects.

For the last two scenarios where the true reference was unknown, one can observe that SKAT-MCN, after Cauchy combination, surpassed both SKAT and the burden test for most cases, except for five-category data with rare variants (Figure 3—homogeneous, Figure

4—heterogeneous). When the genetic effects were homogeneous and the number of categories was small (Figure 3a, common variants; Figure 3c, rare variants), SKAT-MCN consistently surpassed the burden test and SKAT. SKAT suffered from combining the categories that had very distinct genetic effects into a single group, which can lead to substantial power loss. When there were five categories (Figure 3b, common variants; Figure 3d, rare variants), SKAT-MCN was still the most powerful when the number of categories was small, but slightly lost power for five-category data. As we know, the combined p -value returned by Cauchy combination will be larger when nonsignificant individual tests increase. The increasing number of categories and the sparsity introduced by rare variants might challenge SKAT-MCN when the true reference was unknown. Nevertheless, it is uncommon that the number of categories exceeds five in practice. Therefore, SKAT-MCN is expected to be the most powerful under this scenario. For the last scenario (Figure 4), where the true reference was unknown and genetic effects were

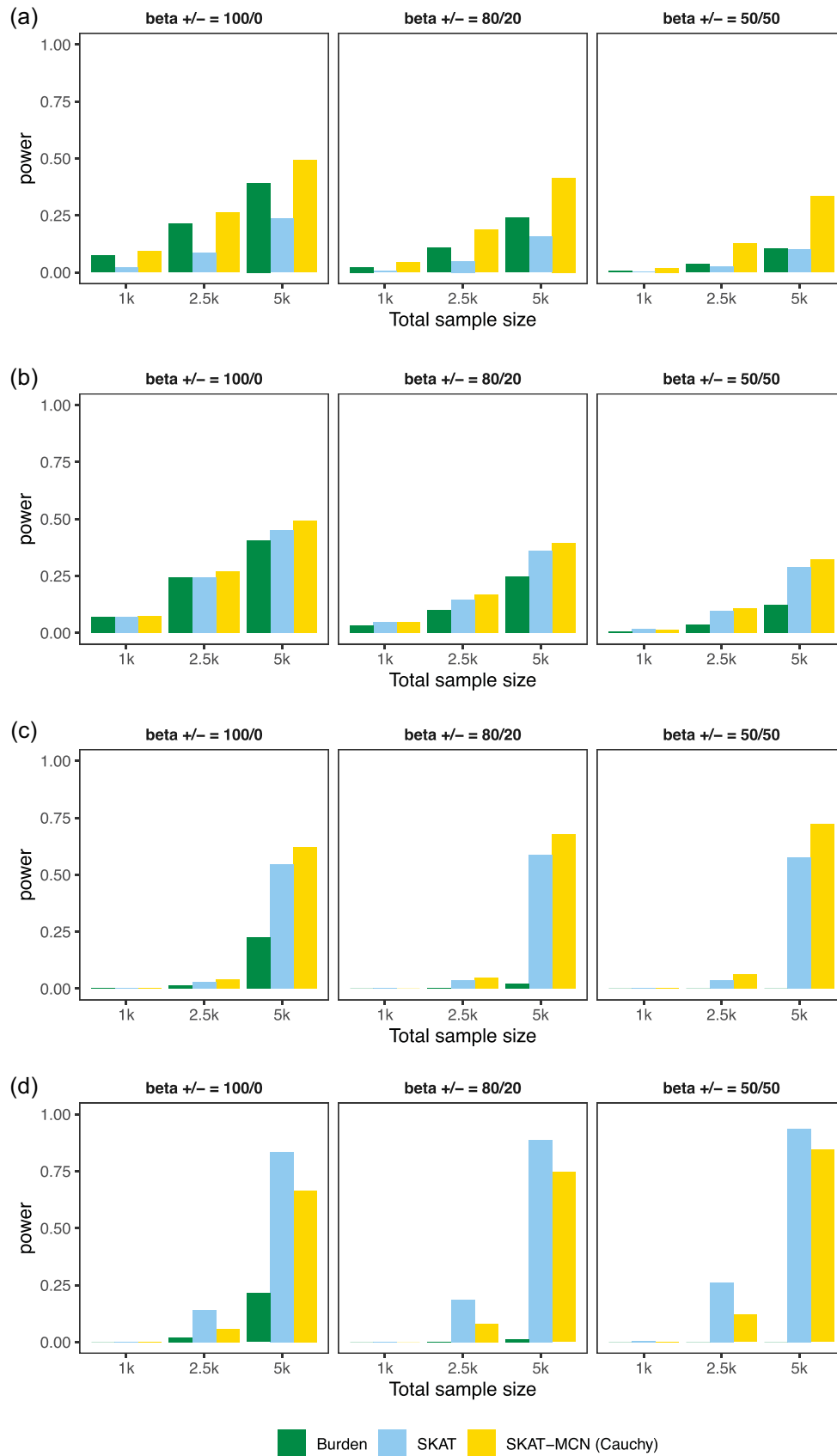


FIGURE 3 (See caption on next page)

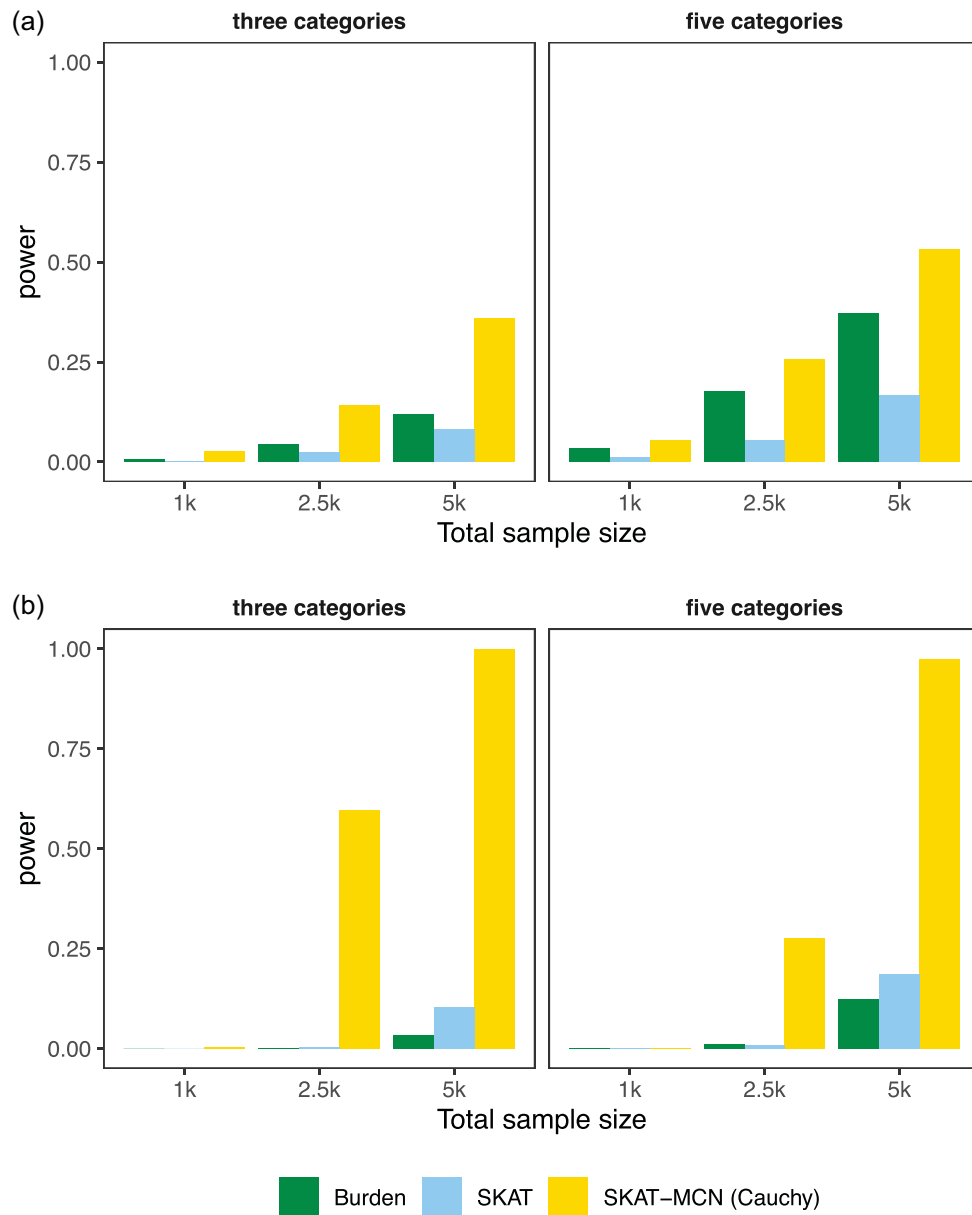


FIGURE 4 Comparisons of statistical power under simulation scenario 4: heterogeneous genetic effects across categories and unknown reference category (for nominal outcomes). (a) Nominal data with three- and five-category outcomes and common variants. (b) Nominal data with three- and five-category outcomes and rare variants. For sequence kernel association test (SKAT)-MCN, we tried each category as the reference in analyses, and then aggregated the individual p -values by Cauchy combination. The unweighted linear kernel was applied to both SKAT-MC and SKAT for the common variants setting, and the weighted linear kernel was applied to SKAT-MC, SKAT, and the burden test for the rare variants setting. The weights were $\text{dbeta}(\text{MAF}, 1, 25)$ for three-category data and $\text{dbeta}(\text{MAF}, 1, 10)$ for five-category data.

FIGURE 3 Comparisons of statistical power under simulation scenario 3: homogeneous genetic effects across categories and unknown reference category (for nominal outcomes). (a) Nominal data with three-category outcomes and common variants. (b) Nominal data with five-category outcomes and common variants. (c) Nominal data with three-category outcomes and rare variants. (d) Nominal data with five-category outcomes and rare variants. For sequence kernel association test (SKAT)-MCN, we tried each category as the reference in analyses, and then aggregated the individual p -values by Cauchy combination. The unweighted linear kernel was applied to both SKAT-MC and SKAT for the common variants setting, and the weighted linear kernel was applied to SKAT-MC, SKAT, and the burden test for the rare variants setting. The weights were $\text{dbeta}(\text{MAF}, 1, 25)$ for three-category data and $\text{dbeta}(\text{MAF}, 1, 10)$ for five-category data.

heterogeneous across categories, SKAT-MCN had substantial power gain compared to SKAT and the burden test for common and rare variants. In short, SKAT-MCN (with Cauchy combination) is tailored to deal with unknown true reference, heterogeneity effects across categories, and negatively associated variants, which explain the superiority.

4.2 | Application to the Polish breast cancer genome-wide association study

We applied SKAT-MC to a population-based breast cancer case-control study conducted in Poland between 2000 and 2003 (Garcia-Closas et al., 2006). The data was provided by OncoArray Consortium (Amos et al., 2017), including 1931 cases that were recently diagnosed with either histologically or cytologically confirmed incident in situ or invasive breast cancer, and 2045 controls that were living in the same place without a history of breast cancer. Tumor characteristics on ER (1076 positive, 557 negative, 298 missing), PR (849 positive, 779 negative, 303 missing) and human epidermal growth factor receptor 2 (HER2) (121 positive, 1004 negative, 806 missing) were documented for each case. The genotype data were derived from OncoArray which was a Illumina genome-wide customer array. Details on genotyping calling, quality control and imputation were described elsewhere (Amos et al., 2017). The genotypes were dosage data imputed to 1000 Genomes Project (Phase3) reference panel (Siva, 2008). We included variants with an imputation score >0.3 , removed rare variants with the MAF <0.05 , and annotated the remaining variants to 23,171 genes on 22 autosomes and X chromosome according to GRCh37. When annotating the variants, we removed pseudo-genes and included the 20 kb upstream region of the transcription start site to each gene. The Bonferroni adjusted genome-wide significance level was $\alpha = 2.16 \times 10^{-6}$ ($\approx 0.05/23171$).

We conducted the genome-wide association analysis by comparing ER+, ER– breast cancer and the control. We first imputed the missing values in ER by logistic regression with covariates PR and HER2 using R package “mice” (van Buuren & Groothuis-Oudshoorn, 2011), ending up with 1282 subjects with ER+, 649 subjects with ER– and 2045 controls. SKAT-MCN was applied by setting each category as the reference to obtain individual p -values and then followed by Cauchy combination to calculate an omnibus SKAT-MCN p -value. SKAT was performed by combining the ER+ and ER– groups into a case group. Both models adjusted for age and the first five principle components of the genotypes to address potential population stratification. We used the unweighted linear kernel for both methods.

Both SKAT and SKAT-MCN detected *FGFR2* as genome-wide significant ($p = 1.79 \times 10^{-7}$ and 5.81×10^{-7} , respectively for SKAT-MCN and SKAT). Moreover, when we inspected the individual SKAT-MCN tests, we found that the p -values reached genome-wide significance using ER+ or the control as the reference ($p = 1.07 \times 10^{-7}$ using ER+ as the reference, $p = 1.35 \times 10^{-7}$ using control as the reference). However, using ER– as the reference, the association was no longer genome-wide significant ($p = 0.028$). These results suggest that the *FGFR2* effect was more dominant on ER+ cancers compared to ER–, as has been shown in studies in Chinese and European women (Chan et al., 2012; Garcia-Closas et al., 2008). Further, SKAT detected *CYP11A* as genome-wide significant ($p = 7.19 \times 10^{-7}$) for breast cancer. Using SKAT-MCN with Cauchy combination, *CYP11A* was close but didn't pass the genome-wide significance threshold ($p = 4.27 \times 10^{-6}$). However, individual SKAT-MCN test with control being the reference generated a significant association ($p = 1.43 \times 10^{-6}$). It indicates that the genetic effects of *CYP11A* were similar on ER– and ER+ cancers, but were different from that on control.

Because of the modest sample size of this study, SKAT-MCN didn't detect more associated genes at the genome-wide significance level. However, some discoveries were worthy of further investigations. *CUPID1* (also called *LINC01488*, $p = 2.48 \times 10^{-5}$ for ER+ as the reference, $p = 0.033$ for ER– as the reference, and $p = 1.97 \times 10^{-4}$ for control as the reference), a long noncoding region, is predominantly expressed in ER+ breast cancer cell lines (Betts et al., 2017), but has no evidence on promoting ER– breast cancer. *KLF11* ($p = 3.59 \times 10^{-3}$ for ER+ as the reference, $p = 3.05 \times 10^{-5}$ for ER– as the reference, $p = 3.57 \times 10^{-3}$ for control as the reference) is likely to influence ER– breast cancer since the Krüppel-like factors (KLFs) have complicated effects on ER-related signaling pathways (J. Zhang, Li, et al., 2020).

4.3 | Application to educational attainment in UK Biobank

To illustrate that SKAT-MC is adequate to large-scale GWAS, we investigated the genetic underpinning for educational attainment using the UK Biobank data. The data was downloaded from Data-Field 6138. We compared three degrees that can be considered as ordinal-O-levels/GCSEs or equivalent, Other professional qualifications (e.g., nursing, teaching) and College or University degree. Subjects with multiple degrees would keep the highest one. We extracted 147, 694 subjects from the original data,

and then removed nonwhite subjects, three-degree related subjects, and subjects with excessive heterozygosity, mismatched sex, and high sex chromosome aneuploidy. These QC metrics were provided by UK Biobank. The final data set contained 127,127 subjects, including 45,155 subjects with O-levels degree, 17,781 subjects with professional degree, and 64,191 subjects with college degree. For variant-level QC, we extracted variants with $\text{INFO} > 0.6$, $0.001 < \text{MAF} < 0.05$, Hardy-Weinberg equilibrium test p -value greater than 10^{-6} , and we removed multiallelic sites. Finally, we partitioned the genome (excluding the sex chromosomes) into 26,528 genes according to GRCh 37 release 13, where the 20kb upstream region was incorporated into each gene. These tested genes span a wide spectrum of functions, including protein-coding genes, long-non-coding RNAs (lncRNA), micro RNAs (miRNA), and other different functional RNAs and segments. We comprehensively compared SKAT-MCN, SKAT-MCO, SKAT, and the burden test (both GLM and POM). For SKAT-MCN, we tried three categories as the reference as well as combined the

individual p -values using the Cauchy combination. For SKAT, we combined the O-level degree category and the professional degree category, that is, we compared college degree or not using SKAT. All SKAT-type tests utilized the weighted linear kernel with weight $\text{dbeta}(\text{MAF}, 1, 25)$. Because the smallest category had more than 10% data, the type I error would be well-controlled by SKAT-MC. The burden test also used $\text{dbeta}(\text{MAF}, 1, 25)$ as weight to linearly combine variants. For all the tests, age, sex, and the first 40 genetic principal components were adjusted.

The significant genes identified by SKAT-MCO after Bonferroni correction ($p < 0.05/26528 = 1.88 \times 10^{-6}$) are present in Table 4 (Supporting Information: Table S3 for all testings results). SKAT-MCO identified 21 genes in the genome, followed by SKAT-MCN (ref. college degree) which identified nine genes, and SKAT-MCN (Cauchy) which identified six genes, whereas burden (POM) identified two genes and SKAT identified four genes. Fifteen genes identified by SKAT-MCO were concentrated on chromosome 3 at the genetic location p21.31, which is a well-known locus associated with

Chr	Gene	Begin	End	Rare variants	p -Value
1	URB2	229741994	229795947	76	1.05×10^{-6}
3	NCKIPSD	48691277	48723348	41	1.13×10^{-6}
3	IP6K2	48705436	48754654	72	1.34×10^{-6}
3	ARIH2OS	48935221	48956818	39	1.84×10^{-6}
3	QRICH1	49047140	49131806	118	1.59×10^{-6}
3	C3orf84	49195067	49229291	39	2.24×10^{-9}
3	IHO1	49215861	49295539	106	8.14×10^{-7}
3	C3orf62	49286029	49314665	32	1.77×10^{-7}
3	RHOA	49376578	49449409	87	6.36×10^{-8}
3	RNF123	49706990	49758962	67	4.68×10^{-7}
3	AMIGO3	49734262	49757117	38	5.36×10^{-8}
3	GMPPB	49737349	49761384	40	6.01×10^{-11}
3	INKA1	49820694	49842463	37	1.50×10^{-7}
3	UBA7	49822642	49851386	45	6.09×10^{-8}
3	MIR5193	49823570	49843678	35	8.61×10^{-8}
3	MST1R	49904435	49941306	35	1.50×10^{-6}
6	TRT-AGT2-2	27632474	27652547	42	1.37×10^{-6}
6	MIR30A	72093254	72113324	41	1.12×10^{-6}
7	MAD1L1	1835431	2272580	1984	2.05×10^{-7}
10	NRAP	115328473	115423800	170	1.12×10^{-6}
13	PCCA	100721347	101182689	796	6.28×10^{-7}

Note: Rare variants represent the number of rare variants ($0.001 < \text{MAF} < 0.05$) included in the analysis for each gene.

TABLE 4 Significant genes identified by sequence kernel association test (SKAT)-MCO that were associated with educational attainment.

intelligence (Hill et al., 2019), educational attainment (Davies et al., 2016; Kichaev et al., 2019; Okbay et al., 2022), and other neurological disorders such as insomnia (Watanabe et al., 2022) and Alzheimer's disease (Kulminski et al., 2022). Compared with a previous study (Davies et al., 2016) that used similar sample size ($N = 112, 151$) but included both common and rare variants ($MAF > 0.001$), 14 out of 21 genes identified by SKAT-MCO were new. Among the new identified genes that were located on other chromosomes, *MAD1L1* on chromosome 7 was associated with schizophrenia (Ikeda et al., 2019; Kulminski et al., 2022); *NRAP* on chromosome 10 was associated with educational attainment (Okbay et al., 2022); and *PCCA* on chromosome 13 was related to educational attainment (J. J. Lee, Wedow, et al., 2018). These results clearly demonstrated the strength of SKAT-MCO on gene-based test for rare variants compared with the traditional burden test. SKAT-MCO also outperformed SKAT-MCN and SKAT by taking advantage of the order information and the heterogeneity among categories.

5 | DISCUSSION

In this paper, we proposed a statistically rigorous and flexible approach (SKAT-MC) to assess the association between a multicategorical outcome (nominal or ordinal) and a set of variants (common or rare) in GWAS. Our method combines the kernel machine regression framework in GWAS with the classic generalized logit model and the proportional odds model for modeling multicategorical outcomes. Because of this, our model can detect association signals under a wide range of association patterns irrespective of whether the genetic effects are homogeneous or heterogeneous across categories (the advantage of generalized logit and proportional odds models compared to collapsing multiple categories into two categories) and whether the multiple genetic variants have different effect sizes and directions (the advantage of the kernel-based approaches). Via extensive simulations, we showed the superior performance of SKAT-MC over potential competitors. Based on the score test and analytical calculation of p -values, SKAT-MC is computationally very fast, suitable for large-scale GWAS.

Two versions of SKAT-MC were proposed, one for ordinal outcomes (SKAT-MCO) and one for nominal outcomes (SKAT-MCN). SKAT-MCO, as far as we know, is the first statistical method that explicitly incorporates the order relationship in set-based analysis for GWAS except for rudimentary methods (such as the burden test). In some sense, an ordinal outcome can be considered as a special case of a nominal outcome with

constraints on the order relationship. In classic statistical literatures, the generalized logit model can be considered as an generalization of the proportional odds model, with more flexible model specification and more parameters to be estimated. From this perspective, SKAT-MCN can also be applied even when the outcome is ordinal—it will still safeguard us from excessive type I errors. However, if the order between categories does exist, ignoring them will lead to a reduction in power, sometimes substantially, as confirmed in our simulation studies.

For SKAT-MCN, the selection of reference category substantially affects model performance, even though the type I error is still well-controlled. Nevertheless, the order in SKAT-MCO ensures that we may only accumulate categories forward or backward. As we have shown, given a particular gene, using the category that reflects the largest disparity among categories will lead to the highest power. This makes intuitive sense because it contrasts the category that is the most different from other categories in the genetic underpinning. However, which category is the most genetically different is unknown, and can differ from gene to gene even for the same outcome. In such a context, we recommend trying each category as the reference, obtaining individual p -values and then aggregating them by Cauchy combination (Y. Liu & Xie, 2020) for an overall association assessment. Through this “omnibus” test, our method is robust against mis-specification of references. If the primary goal is to examine overall association for each gene with a multicategorical outcome, it is advisable to use the “omnibus” test. On the other hand, if the interest is to verify the association between a particular gene and a subtype of disease, such as in candidate gene analysis, using the target subtype as the reference is adequate. A significant result (if exists) indicates a genetic disparity between this subtype and others. In addition to merely association testing, comparing the individual p -values from tests using different references may suggest the underlying genetic pattern of association. As an example, by inspecting the p -values generated by using ER+, ER– and control as the reference for analyzing *FGFR2*, one can infer the genetic effect on ER+ is much stronger than that on ER–. This preliminary insight is particularly useful for researchers to further investigate the pathological effect imposed by *FGFR2*.

For rare variants association testing, it is a common practice to give higher weights to rarer variants (Wu et al., 2011). However, it should be cautious to assign weights to rare variants when analyzing multicategorical data. Although rarer variants may play a more important role in the disease pathology, the test will be driven by them if assigning extremely high weights. Then, the

regularization assumptions for the asymptotic distribution of test statistics under the null will be violated. Consequently, the type I error cannot be controlled at the extreme tail, such as 1.0×10^{-6} . This problem does not only apply to SKAT-MC, but also apply to the burden score or Wald test. And the problem will deteriorate when samples are severely unbalanced across categories. In the SKAT paper, the authors proposed $\text{dbeta}(\text{MAF}, 1, 25)$ as the weight for continuous and binary outcomes, which is sometimes too extreme for multicategorical outcomes, if for example, Categories 1 and 2 each has 5% samples and Category 3 has 90% samples. Thus we propose to use $\text{dbeta}(\text{MAF}, 1, 10)$ as weights in the weighted linear kernel in the above extreme case. Users are also encouraged to do sensitivity analysis for different weights to inspect the robustness.

Potential extensions for SKAT-MC are in three folds. First, the strategy of selecting rare variants into the variant set is not limited to including all variants in a gene, as we can refer to variant coding annotation (X. Li et al., 2020; Wang et al., 2021) or noncoding annotation (Z. Li, Li, et al., 2022) categories to select variants. Alternatively, there are nongene-centric approaches by grouping rare variants using agnostic windows with fixed lengths (X. Li et al., 2020; Morrison et al., 2017) or dynamic lengths (Z. Li, Li, et al., 2022; Z. Li et al., 2019). Second, besides assigning weights based on minor allele frequencies, the weights for rare variants can be assigned by incorporating functional annotations. The commonly used annotation weights include annotation principal components (X. Li et al., 2020; Zhou et al., 2023), CADD (Kircher et al., 2014; Rentzsch et al., 2019), MACIE (X. Li, Yung, et al., 2022), among others (Gaynor et al., 2022; P. H. Lee, Lee, et al., 2018). Last, using summary statistics in SKAT-MC. For privacy reasons, acquiring individual-level genetic data to do inference is difficult; also sharing and storing large-scale genetic data is a common obstacle for modern GWAS. Therefore, summary statistics-based methods are increasingly appealing (X. Li, Quick, et al., 2022; D. J. Liu et al., 2014), because it not only alleviates the sharing and storage burden, but it also makes meta-analysis possible across multiple studies and ethnicity groups.

Identification and characterization of the genetic heterogeneity among disease subtypes is an indispensable step toward a full understanding of the disease etiology and a better strategy for disease prevention and management. We believe that SKAT-MC can be a useful tool for our effort in such a direction.

ACKNOWLEDGMENTS

The authors thank editors and reviewers for their insightful comments which have led to a significant improvement of this article. Zhan's research was

supported in part by the National Key R&D Program of China (grant number 2022YFA1305400). Zhao's research was supported in part by the National Institute of Health (grant number U24OD023382).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The National Cancer Institute Polish Breast Cancer Study is publicly available via dbGaP (www.ncbi.nlm.nih.gov/gap; accession number phs001265.v1.p1). The UK Biobank data can be accessed at <https://www.ukbiobank.ac.uk> upon application. The educational attainment data is in Data-Field 6138. R package SKAT-MC can be accessed at <https://github.com/Zhiwen-Owen-Jiang/SKATMC>. The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Zhiwen Jiang  <http://orcid.org/0000-0001-8841-618X>

Ni Zhao  <http://orcid.org/0000-0002-7762-3949>

REFERENCES

- Amos, C. I., Dennis, J., Wang, Z., Byun, J., Schumacher, F. R., Gayther, S. A., Casey, G., Hunter, D. J., Sellers, T. A., Gruber, S. B., Dunning, A. M., Michailidou, K., Fachal, L., Doheny, K., Spurdle, A. B., Li, Y., Xiao, X., Romm, J., Pugh, E., ... Easton, D. F. (2017). The OncoArray consortium: A network for understanding the genetic architecture of common cancers. *Cancer Epidemiology and Prevention Biomarkers*, 26(1), 126–135.
- Bareche, Y., Venet, D., Ignatiadis, M., Aftimos, P., Piccart, M., Rothe, F., & Sotiriou, C. (2018). Unravelling triple-negative breast cancer molecular heterogeneity using an integrative multiomic analysis. *Annals of Oncology*, 29(4), 895–902.
- Betts, J. A., Marjaneh, M. M., Al-Ejeh, F., Lim, Y. C., Shi, W., Sivakumaran, H., Tropée, R., Patch, A.-M., Clark, M. B., & Bartonicek, N. (2017). Long noncoding RNAs CUPID1 and CUPID2 mediate breast cancer risk at 11q13 by modulating the response to DNA damage. *The American Journal of Human Genetics*, 101(2), 255–266.
- Bocher, O., Marenne, G., Saint Pierre, A., Ludwig, T. E., Guey, S., Tournier-Lasserre, E., Perdry, H., & Génin, E. (2019). Rare variant association testing for multicategory phenotype. *Genetic Epidemiology*, 43(6), 646–656.
- Bocher, O., Marenne, G., Tournier-Lasserre, E., Génin, E., Perdry, H., & Consortium, F. (2021). Extension of SKAT to multi-category phenotypes through a geometrical interpretation. *European Journal of Human Genetics*, 29(5), 736–744.
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multi-variate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1–67.
- Chan, M., Ji, S., Liaw, C., Yap, Y., Law, H., Yoon, C., Wong, C., Yong, W., Wong, N., & Ng, R. (2012). Association of common

- genetic variants with breast cancer risk and clinicopathological characteristics in a Chinese population. *Breast Cancer Research and Treatment*, 136(1), 209–220.
- Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines: And other kernel-based learning methods*. Cambridge University Press.
- Davenport, C. A., Maity, A., Sullivan, P. F., & Tzeng, J. Y. (2018). A powerful test for SNP effects on multivariate binary outcomes using kernel machine regression. *Statistics in Biosciences*, 10(1), 117–138.
- Davies, G., Marioni, R. E., Liewald, D. C., Hill, W. D., Hagenaars, S. P., Harris, S. E., Ritchie, S. J., Luciano, M., Fawns-Ritchie, C., Lyall, D., Cullen, B., Cox, S. R., Hayward, C., Porteous, D. J., Evans, J., McIntosh, A. M., Gallacher, J., Craddock, N., Pell, J. P., ... Deary, I. J. (2016). Genome-wide association study of cognitive functions and educational attainment in UK Biobank (n=112 151). *Molecular Psychiatry*, 21(6), 758–767.
- Garcia-Closas, M., Brinton, L. A., Lissowska, J., Chatterjee, N., Peplonska, B., Anderson, W. F., Szeszenia-Dabrowska, N., Bardin-Mikolajczak, A., Zatonski, W., Blair, A., Kalaylioglu, Z., Rymkiewicz, G., Mazepa-Sikora, D., Kordek, R., Lukaszek, S., & Sherman, M. E. (2006). Established breast cancer risk factors by clinically important tumour characteristics. *British Journal of Cancer*, 95(1), 123–129.
- Garcia-Closas, M., Hall, P., Nevanlinna, H., Pooley, K., Morrison, J., Richesson, D. A., Bojesen, S. E., Nordestgaard, B. G., Axelsson, C. K., Arias, J. I., Milne, R. L., Ribas, G., González-Neira, A., Benítez, J., Zamora, P., Brauch, H., Justenhoven, C., Hamann, U., Ko, Y. D., ... Pharoah, P. D. (2008). Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics. *PLoS Genet*, 4(4), e1000054.
- Gaynor, S. M., Westerman, K. E., Ackovic, L. L., Li, X., Li, Z., Manning, A. K., Philippakis, A., & Lin, X. (2022). STAAR workflow: A cloud-based workflow for scalable and reproducible rare variant analysis. *Bioinformatics*, 38(11), 3116–3117.
- He, Q., Liu, Y., Liu, M., Wu, M., & Hsu, L. (2021). Random-effect based test for multinomial logistic regression: Choice of the reference level and its impact on the testing. medRxiv.
- Hill, W. D., Marioni, R. E., Maghzian, O., Ritchie, S. J., Hagenaars, S. P., McIntosh, A., Gale, C. R., Davies, G., & Deary, I. J. (2019). A combined analysis of genetically correlated traits identifies 187 loci and a role for neurogenesis and myelination in intelligence. *Molecular Psychiatry*, 24(2), 169–181.
- Ikeda, M., Takahashi, A., Kamatani, Y., Momozawa, Y., Saito, T., Kondo, K., Shimasaki, A., Kawase, K., Sakusabe, T., Iwayama, Y., Toyota, T., Wakuda, T., Kikuchi, M., Kanahara, N., Yamamori, H., Yasuda, Y., Watanabe, Y., Hoya, S., Aleksic, B., ... Iwata, N. (2019). Genome-wide association study detected novel susceptibility genes for schizophrenia and shared trans-populations/diseases genetic effect. *Schizophrenia Bulletin*, 45(4), 824–834.
- Kichaev, G., Bhatia, G., Loh, P.-R., Gazal, S., Burch, K., Freund, M. K., Schoech, A., Pasaniuc, B., & Price, A. L. (2019). Leveraging polygenic functional enrichment to improve GWAS power. *The American Journal of Human Genetics*, 104(1), 65–75.
- Kircher, M., Witten, D. M., Jain, P., O’Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3), 310–315.
- Kulminski, A. M., Loiko, E., Loika, Y., & Culminkaya, I. (2022). Pleiotropic predisposition to Alzheimer’s disease and educational attainment: Insights from the summary statistics analysis. *GeroScience*, 44(1), 265–280.
- Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., Nguyen-Viet, T. A., Bowers, P., Sidorenko, J., Karlsson Linnér, K. R., Fontana, M. A., Kundu, T., Lee, C., Li, H., Li, R., Royer, R., Timshel, P. N., Walters, R. K., Willoughby, E. A., ... Cesarini, D. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics*, 50(8), 1112–1121.
- Lee, P. H., Lee, C., Li, X., Wee, B., Dwivedi, T., & Daly, M. (2018). Principles and methods of in-silico prioritization of non-coding regulatory variants. *Human Genetics*, 137, 15–30.
- Lee, S., Abecasis, G. R., Boehnke, M., & Lin, X. (2014). Rare-variant association analysis: Study designs and statistical tests. *The American Journal of Human Genetics*, 95(1), 5–23.
- Lee, S., Wu, M. C., & Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13(4), 762–775.
- Li, B., & Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *The American Journal of Human Genetics*, 83(3), 311–321.
- Li, X., Li, Z., Zhou, H., Gaynor, S. M., Liu, Y., Chen, H., Sun, R., Dey, R., Arnett, D. K., Aslibekyan, S., Ballantyne, C. M., Bielak L. F., Blangero, J., Boerwinkle, E., Bowden, D. W., Broome, J. G., Conomos, M. P., Correa, A., Adrienne Cupples, L., ... Lin, X. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nature Genetics*, 52(9), 969–983.
- Li, X., Quick, C., Zhou, H., Gaynor, S. M., Liu, Y., Chen, H., Selvaraj, M. S., Sun, R., Dey, R., Arnett, D. K., Bielak, L. F., Bis, J. C., Blangero, J., Boerwinkle, E., Bowden, D. W., Brody, J. A., Cade, B. E., Correa, A., Cupples, L. A., ... Lin, X. (2022). Powerful, scalable and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies. *Nature Genetics*, 55, 154–164.
- Li, X., Yung, G., Zhou, H., Sun, R., Li, Z., Hou, K., Zhang, M. J., Liu, Y., Arapoglou, T., Wang, C. Ionita-Laza, I. & Lin, X. (2022). A multi-dimensional integrative scoring framework for predicting functional variants in the human genome. *The American Journal of Human Genetics*, 109(3), 446–456.
- Li, Z., Li, X., Liu, Y., Shen, J., Chen, H., Zhou, H., Morrison, A. C., Boerwinkle, E., & Lin, X. (2019). Dynamic scan procedure for detecting rare-variant association regions in whole-genome sequencing studies. *The American Journal of Human Genetics*, 104(5), 802–814.
- Li, Z., Li, X., Zhou, H., Gaynor, S. M., Selvaraj, M. S., Arapoglou, T., Quick, C., Liu, Y., Chen, H., Sun, R., Dey, R., Arnett, D. K., Auer, P. L., Bielak, L. F., Bis, J. C., Blackwell, T. W., Blangero, J., Boerwinkle, E., Bowden, D. W., ... Lin, X. (2022). A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nature Methods*, 19, 1599–1611.

- Liang, H., Yang, X., Chen, L., Li, H., Zhu, A., Sun, M., Wang, H., & Li, M. (2015). Heterogeneity of breast cancer associations with common genetic variants in FGFR2 according to the intrinsic subtypes in Southern Han Chinese women. *Biomed Research International*, 2015, 626948.
- Liu, D., Lin, X., & Ghosh, D. (2007). Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics*, 63(4), 1079–1088.
- Liu, D. J., Peloso, G. M., Zhan, X., Holmen, O. L., Zawistowski, M., Feng, S., Nikpay, M., Auer, P. L., Goel, A., Zhang, H., Peters, U., Farrall, M., Orho-Melander, M., Kooperberg, C., McPherson, R., Watkins, H., Willer, C. J., Hveem, K., Melander, O., ... Abecasis, G. R. (2014). Meta-analysis of gene-level tests for rare variant association. *Nature Genetics*, 46(2), 200–204.
- Liu, H., Tang, Y., & Zhang, H. H. (2009). A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics & Data Analysis*, 53(4), 853–856.
- Liu, M., Liu, Y., Wu, M. C., Hsu, L., & He, Q. (2021). A method for subtype analysis with somatic mutations. *Bioinformatics*, 37(1), 50–56.
- Liu, Y., & Xie, J. (2020). Cauchy combination test: A powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115(529), 393–402.
- MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., Pendlington, Z. M., Welter, D., Burdett, T., Hindorf, L., Flicek, P., Cunningham, F., & Parkinson, H. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, 45(D1), D896–D901.
- Madsen, B. E., & Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics*, 5(2), e1000384.
- Morrison, A. C., Huang, Z., Yu, B., Metcalf, G., Liu, X., Ballantyne, C., Coresh, J., Yu, F., Muzny, D., Feofanova, E., Rustagi, N., Gibbs, R., & Boerwinkle, E. (2017). Practical approaches for whole-genome sequence analysis of heart-and blood-related traits. *The American Journal of Human Genetics*, 100(2), 205–215.
- Okbay, A., Wu, Y., Wang, N., Jayashankar, H., Bennett, M., Nehzati, S. M., Sidorenko, J., Kweon, H., Goldman, G., Gjorgjieva, T., Jiang, Y., Hicks, B., Tian, C., Hinds, D. A., Ahlskog, R., Magnusson, P. K. E., Oskarsson, S., Hayward, C., Campbell, A., ... Young, A. I. (2022). Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nature Genetics*, 54(4), 437–449.
- Prat, A., Pineda, E., Adamo, B., Galvan, P., Fernandez, A., Gaba, L., Diez, M., Viladot, M., Arance, A., & Munoz, M. (2015). Clinical implications of the intrinsic molecular subtypes of breast cancer. *Breast*, 24(Suppl. 2), S26–S35.
- Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, 47(D1), D886–D894.
- Siva, N. (2008). 1000 genomes project. *Nature Biotechnology*, 26(3), 256–257.
- Su, Z., Marchini, J., & Donnelly, P. (2011). HAPGEN2: Simulation of multiple disease SNPs. *Bioinformatics*, 27(16), 2304–2305.
- Wang, Q., Dhindsa, R. S., Carss, K., Harper, A. R., Nag, A., Tachmazidou, I., Vitsios, D., Deevi, S. V., Mackay, A., Muthas, D., Hühn, M., Monkley, S., Olsson, H., AstraZeneca Genomics Initiative, Wasilewski, S., Smith, K. R., March, R., Platt, A., Haefliger, C., & Petrovski, S. (2021). Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature*, 597(7877), 527–532.
- Watanabe, K., Jansen, P. R., Savage, J. E., Nandakumar, P., Wang, X., Hinds, D. A., Gelernter, J., Levey, D. F., Polimanti, R., Stein, M. B., Van Someren, E. J. W., Smit, A. B., & Posthuma, D. (2022). Genome-wide meta-analysis of insomnia prioritizes genes associated with metabolic and psychiatric pathways. *Nature Genetics*, 54(8), 1125–1132.
- Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1), 82–93.
- Yue, J.-X., & Liti, G. (2019). simuG: A general-purpose genome simulator. *Bioinformatics*, 35(21), 4442–4444.
- Zhang, H., Ahearn, T. U., Lecarpentier, J., Barnes, D., Beesley, J., Qi, G., Jiang, X., O'Mara, T. A., Zhao, N., Bolla, M. K., Dunning, A. M., Dennis, J., Wang, Q., Abu Ful Z., Aittomäki K., Andrulis, I. L., Anton-Culver, Volker Arndt, H., Aronson, K. J., ... García-Closas, M. (2020). Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nature Genetics*, 52(6), 572–581.
- Zhang, H., Zhao, N., Ahearn, T. U., Wheeler, W., García-Closas, M., & Chatterjee, N. (2021). A mixed-model approach for powerful testing of genetic associations with cancer risk incorporating tumor characteristics. *Biostatistics*, 22(4), 772–788.
- Zhang, J., Li, G., Feng, L., Lu, H., & Wang, X. (2020). Krüppel-like factors in breast cancer: Function, regulation and clinical relevance. *Biomedicine & Pharmacotherapy*, 123, 109778.
- Zhou, H., Arapoglou, T., Li, X., Li, Z., Zheng, X., Moore, J., Asok, A., Kumar, S., Blue, E. E., Buyske, S., Cox, N., Felsenfeld, A., Gerstein, M., Kenny, E., Li, B., Matise, T., Philippakis, A., Rehm, H. L., Sofia, H. J., ... Lin, X. (2023). FAVOR: functional annotation of variants online resource and annotator for variation across the human genome. *Nucleic Acids Research*, 51(D1), D1300–D1311.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Jiang, Z., Zhang, H., Ahearn, T. U., Garcia-Closas, M., Chatterjee, N., Zhu, H., Zhan, X., & Zhao, N. (2023). The sequence kernel association test for multicategorical outcomes. *Genetic Epidemiology*, 1–18. <https://doi.org/10.1002/gepi.22527>