AUTHOR ACCEPTED MANUSCRIPT VERSION OF THE PAPER BY JOHNSEN, BAKKE, BJØRNLAND, DEWAN AND LANGAAS IN STATISTICS IN MEDICINE (2023) DOI: 10.1002/SIM.9746

Saddlepoint approximations to score test statistics in logistic regression for analysing genome-wide association studies

Pål V. Johnsen[1,2] | Øyvind Bakke[2] | Thea Bjørnland[2] | Andrew Thomas DeWan[3] | Mette Langaas[2]

[1]SINTEF DIGITAL, Oslo, Norway
[2]Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway
[3]Department of Chronic Disease Epidemiology and Center for Perinatal, Pediatric and Environmental Epidemiology, Yale School of Public Health, New Haven, Connecticut, USA

Correspondence
Pål V. Johnsen,
Email: pal.johnsen@sintef.no

Summary

We investigate saddlepoint approximations of tail probabilities of the score test statistic in logistic regression for genome-wide association studies. The inaccuracy in the normal approximation of the score test statistic increases with increasing imbalance in the response and with decreasing minor allele counts. Applying saddlepoint approximation methods greatly improves the accuracy, even far out in the tails of the distribution. By using exact results for a simple logistic regression model, as well as simulations for models with nuisance parameters, we compare double saddlepoint methods for computing two-sided p-values and mid-p-values. These methods are also compared to a recent single saddlepoint procedure. We investigate the methods further on data from UK Biobank with skin and soft tissue infections as phenotype, using both common and rare variants.

KEYWORDS:
GWAS, imbalanced binary response, logistic regression, score test, saddlepoint approximation

1 | INTRODUCTION

In genome-wide association studies (GWAS), each single nucleotide polymorphism (SNP) is tested individually for association with a particular phenotype. If the phenotype is binary (e.g. disease / not disease) a logistic regression model may be used to model the probability of disease, and our interest lies in testing one SNP at a time against a specified null model. In a modern biobank including several hundred thousands of SNPs, rejection of the null hypothesis for each individual SNP needs to be evaluated at a low p-value threshold, typically $5 \cdot 10^{-8}$, in order to control the family-wise error rate.[1,2] Tests applied under an assumption of normality may yield inflated error rates if the response variable in a logistic regression model is imbalanced. The severity in this flaw increases with decreasing minor allele frequencies (MAF).[3,4] Hence there is a need to develop tests for logistic regression models where the response and covariate of interest are imbalanced. Since the null model is the same for all SNPs in a GWAS, the score test - which only requires estimation of parameters under the null - can be computationally advantageous.

As an example, we consider a follow-up study on skin and soft tissue infection (SSTI) using UK-biobank data, motivated by Rogne et al.[5] Using data on unrelated white European individuals with no prior history of SSTI at recruitment, we obtain 6.5 years of follow-up data on approximately 300 000 individuals, out of which approximately 0.7% were diagnosed with SSTI during follow-up, and classified as cases. The overall sample size is large, but there are relatively few cases. Relying on asymptotic normality of the score test statistic may therefore yield spurious results. Both Ma et al.[3] and Dey et al.[4] have illustrated the

inaccuracy of the normal approximation for logistic regression models in GWAS. A solution proposed by Ma et al.[3] is to apply the Firth[6] bias-corrected logistic regression test. As Firth's test was found to be computationally slow in genome-wide testing, a test based on a saddlepoint approximation to a score statistic was proposed by Dey et al.[4] Saddlepoint methods to compute p-values in logistic regression have previously been developed for example for small sample inference[7] and evaluated against some specific exact conditional one-sided tests[8]. Although GWAS sample sizes in constrast are large, similar issues with the assumption of normality may arise when the response and covariate of interest are both highly imbalanced.[3]

Our theoretical contribution to the ongoing development of score tests for genome-wide association studies also concerns the use of saddlepoint approximations. First, in Section 2 we establish the discrete and bounded nature of the score, and derive the exact conditional distribution of the score test statistic for particular examples of logistic regression models, the simplest being a model with intercept and genetic variant only. Second, in Section 3 we propose a continuity-corrected double saddlepoint approximation to conditional tail probabilities of the score statistic for models with more complex covariate patterns. Due to the discrete nature of the proposed tests, we also consider saddlepoint methods for computing mid-p-values.[9] We then show that a single saddlepoint method for computing mid-p-values based on the efficient score, or equivalently a null-orthogonal reparameterization of the logistic regression model, coincides with the SPA-test by Dey et al[4] for discrete genetic covariates. A continuity-correction to the SPA-test then follows naturally, and we compare this approach to the double saddlepoint method. On simulated data where an exact test is available, we study in Section 4 the type I error rates of saddlepoint methods, as compared to the normal assumption and the exact test. We also study the performance of saddlepoint approximation methods on simulated data with more complex covariate patterns. Finally, we apply these methods in the follow-up study of SSTIs using UK-biobank data in Section 5.

2 | THE SCORE TEST

2.1 | Notation, statistical model and hypotheses

We consider tests for genotype–phenotype associations in large cohorts or populations. We assume that binary phenotypes, Y_i, non-genetic covariates $\boldsymbol{x}_i$ and allele counts g_i for a single variant, $i = 1, \ldots, n$, have been collected from n individuals. We consider biallelic counts in which $g_i = 0$, 1 or 2. We model the relationship between the response and the covariates in a logistic regression model in which the Y_i are independent and Bernoulli distributed with success probability μ_i and

$$\text{logit}\, \mu_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \gamma g_i, \tag{1}$$

$i = 1, \ldots, n$. Here, $\boldsymbol{x}_i$ is a vector of dimension d containing the constant 1 (corresponding to an intercept) and observed values for $d-1$ covariates, $\boldsymbol{\beta}$ is a d-dimensional vector of nuisance parameters and γ is the parameter of interest. Our aim is to perform the hypothesis test

$$H_0 : \gamma = 0 \quad \text{against} \quad H_1 : \gamma \neq 0. \tag{2}$$

In a GWAS, the test is performed for each individual genetic variant at a time. To control the family-wise error rate at 5%, a significance level of $5 \cdot 10^{-8}$ is commonly used for each test.[1]

2.2 | The score test statistic

The score vector is the gradient of the log-likelihood function with respect to the parameters, which for the logistic regression model (1) is

$$\boldsymbol{U} = \begin{pmatrix} \boldsymbol{U}_\beta \\ U_\gamma \end{pmatrix} = \begin{pmatrix} X^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{\mu}) \\ \boldsymbol{g}^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{\mu}) \end{pmatrix}, \tag{3}$$

where $\boldsymbol{Y}$ and $\boldsymbol{g}$ are column vectors of length n with Y_i and g_i as elements respectively, $\boldsymbol{\mu} = E\boldsymbol{Y}$, and X is an $n \times d$ matrix with $\boldsymbol{x}_i^{\mathrm{T}}$ as rows. We have partitioned the score vector according to the parameter of interest, γ, and the nuisance parameters, $\boldsymbol{\beta}$. The score vector has mean $\boldsymbol{0}$ and covariance matrix $\mathrm{E}(\boldsymbol{U}^T\boldsymbol{U}) = F$, by definition referred to as the expected Fisher information. Here,

$$F = \begin{pmatrix} F_{\beta\beta} & \boldsymbol{F}_{\gamma\beta}^{\mathrm{T}} \\ \boldsymbol{F}_{\gamma\beta} & F_{\gamma\gamma} \end{pmatrix} = \begin{pmatrix} X^{\mathrm{T}} W X & X^{\mathrm{T}} W \boldsymbol{g} \\ \boldsymbol{g}^{\mathrm{T}} W X & \boldsymbol{g}^{\mathrm{T}} W \boldsymbol{g} \end{pmatrix}, \tag{4}$$

where W is a diagonal matrix with $\mu_i(1-\mu_i)$ as the ii entry. Using the score test, the null hypothesis of (2) is rejected if there is sufficient distance between the null value $\gamma = 0$ and the maximum likelihood estimate of γ under the alternative hypothesis. To

judge this distance, without actually calculating the estimate of γ under the alternative hypothesis, one uses the partial derivative U_γ of the log-likelihood with respect to γ at $\gamma = 0$, along with the probability distribution of U_γ under the null.

2.2.1 | The normal approximation

Asymptotically, $\boldsymbol{U} \sim \text{MVN}(\boldsymbol{0}, F)$, and in many applications one may approximate the distribution of the score vector $\boldsymbol{U}$ by a multivariate normal distribution. Under the null hypothesis, maximum likelihood estimates $\hat{\boldsymbol{\beta}}$ of the nuisance parameters are obtained by solving $\boldsymbol{U}_\beta = \boldsymbol{0}$ and the conditional distribution of U_γ given $\boldsymbol{U}_\beta = \boldsymbol{0}$ is asymptotically a normal distribution with mean 0 and variance

$$\tilde{F}_{\gamma\gamma} = \boldsymbol{g}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{g} - \boldsymbol{g}^{\mathrm{T}}\boldsymbol{W}X(X^{\mathrm{T}}\boldsymbol{W}X)^{-1}X^{\mathrm{T}}\boldsymbol{W}\boldsymbol{g}. \tag{5}$$

A score test can be performed by comparing the observed test statistic $u = \boldsymbol{g}^{\mathrm{T}}(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$, where $\hat{\boldsymbol{\mu}}$ satisfies logit $\hat{\mu}_i = \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}$, to the univariate normal distribution with mean 0 and variance $\boldsymbol{g}^{\mathrm{T}}\hat{\boldsymbol{W}}\boldsymbol{g} - \boldsymbol{g}^{\mathrm{T}}\hat{\boldsymbol{W}}X(X^{\mathrm{T}}\hat{\boldsymbol{W}}X)^{-1}X^{\mathrm{T}}\hat{\boldsymbol{W}}\boldsymbol{g}$. Of relevance to the next sections, Lin (2005) showed that the score test in GWAS can be expressed in terms of the efficient score [10,11], here denoted by $\tilde{U}_\gamma$, and for model (1) defined by

$$\tilde{U}_\gamma = U_\gamma - \boldsymbol{F}_{\gamma\beta}F_{\beta\beta}^{-1}\boldsymbol{U}_\beta. \tag{6}$$

As noted by Bickel et al. [11] (p. 30) the efficient score may be interpreted in general as the score corresponding to a reparameterization $(\boldsymbol{\beta}, \gamma) \to (\boldsymbol{\alpha}, \gamma)$, by letting $\boldsymbol{\beta}(\boldsymbol{\alpha}, \gamma) = \boldsymbol{\alpha} - F_{\beta\beta}^{-1}\boldsymbol{F}_{\gamma\beta}^{T}\gamma$. With this reparameterization the logistic regression model can be expressed as $\text{logit}(\mu_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\alpha} + \gamma\tilde{g}_i$, with $\tilde{\boldsymbol{g}} = \boldsymbol{g} - X(X^{T}\boldsymbol{W}X)^{-1}X^{T}\boldsymbol{W}\boldsymbol{g}$ the vector of all components $\tilde{g}_i$, and defined as in Dey et al[4]. Let $\tilde{F}$ denote the expected Fisher information of $\tilde{\boldsymbol{U}} = \left(\tilde{\boldsymbol{U}}_{\boldsymbol{\alpha}}^{\mathrm{T}}\ \tilde{U}_\gamma\right)^{\mathrm{T}}$, the reparameterized score vector. The parameter γ and the nuisance parameters $\boldsymbol{\alpha}$ are locally information orthogonal at $\gamma = 0$, which means that $\tilde{\boldsymbol{F}}_{\boldsymbol{\alpha}\gamma}$ and $\tilde{\boldsymbol{F}}_{\gamma\boldsymbol{\alpha}}$ in the expected Fisher information $\tilde{F}$ are zero-vectors (see e.g. Lindsey [12] p. 237). Asymptotically, $\tilde{\boldsymbol{U}}$ is multivariate normal, and so will $\tilde{\boldsymbol{U}}_{\boldsymbol{\alpha}}$ and $\tilde{U}_\gamma$ (univariate) be. As covariance equal to zero ($\tilde{\boldsymbol{F}}_{\gamma\boldsymbol{\alpha}} = 0$) for two normal distributed random variables implies independence, this means that the distribution of $\tilde{U}_\gamma$ conditional on $\tilde{\boldsymbol{U}}_{\boldsymbol{\alpha}} = \boldsymbol{0}$ is asymptotically the same as the unconditional distribution of $\tilde{U}_\gamma$ when the null hypothesis is true with $\hat{\boldsymbol{\mu}}$ treated as a plug-in constant for $\boldsymbol{\mu}$. As $\tilde{U}_\gamma$ is asymptotically $N(0, \tilde{F}_{\gamma\gamma})$ and the observed efficient score $\tilde{u}$ from (6) equals the observed score u from (3), *unconditional* inference on $\tilde{U}_\gamma(\hat{\boldsymbol{\mu}})$ with $\hat{\boldsymbol{\mu}}$ treated as a plug-in constant, leads to the same asymptotic test as *conditional* inference on U_γ given $\boldsymbol{U}_\beta = \boldsymbol{0}$.

2.2.2 | Discrete test statistics

As outlined in Section 1, the normal approximation to the score vector may lead to spurious results for genotype–phenotype associations in logistic regression. Even in large samples the normal approximation may be inaccurate if the sample contains few individuals with response $y_i = 1$ (e.g., having the disease under study) and genotype $g_i > 0$ (carrying the minor allele) since the score U_γ will be discrete, skewed and bounded. When $g_i \in (0, 1, 2)$, we note that $\boldsymbol{g}^{\mathrm{T}}\boldsymbol{Y}$ is an integer and $\boldsymbol{g}^{\mathrm{T}}\boldsymbol{\mu}$ a constant, so that $U_\gamma = \boldsymbol{g}^{\mathrm{T}}\boldsymbol{Y} - \boldsymbol{g}^{\mathrm{T}}\boldsymbol{\mu}$ has support on a subset of a lattice with step 1. The minimum is obtained for $\boldsymbol{Y} = \boldsymbol{0}$ and the maximum for $\boldsymbol{Y} = \boldsymbol{1}$ (a vector of ones), which leads to the following observation:

Observation 1. When $g_i \in \{0, 1, 2\}$, the score U_γ with respect to γ is a bounded lattice random variable with support on $-\boldsymbol{g}^{\mathrm{T}}\boldsymbol{\mu}$, $1 - \boldsymbol{g}^{\mathrm{T}}\boldsymbol{\mu}$, $2 - \boldsymbol{g}^{\mathrm{T}}\boldsymbol{\mu}$, …, $\boldsymbol{g}^{\mathrm{T}}\boldsymbol{1} - \boldsymbol{g}^{\mathrm{T}}\boldsymbol{\mu}$.

The distribution of U_γ given $\boldsymbol{U}_\beta = \boldsymbol{0}$ will be a lattice distribution with a narrower support than described in Observation 1 (see Appendix A). The discrete and potentially skewed nature of the conditional score test might best be illustrated by the construction of the following exact tests for simple logistic regression models. Proofs are given in Appendix B.

Observation 2. Consider a logistic regression model as in (1), but with logit $\mu_i = \beta_0 + \gamma g_i$, henceforth denoted the *intercept model*. Let n_j be the number of individuals with genotype $g_i = j$, $j = 0, 1, 2$, and let logit $\mu = \beta_0$ under the null. Then, the null distribution of U_γ given $U_{\beta_0} = 0$ is a sum of trivariate hypergeometric point probabilities,

$$P(U_\gamma = u \mid U_{\beta_0} = 0) = \sum_{(v_0,v_1,v_2)\in S} \frac{\binom{n_0}{v_0}\binom{n_1}{v_1}\binom{n_2}{v_2}}{\binom{n}{n\mu}} = \sum_{k=\max(\lceil (u^*-n_1)/2\rceil, 0)}^{\min(\lfloor u^*/2\rfloor, n_2)} \frac{\binom{n_0}{n\mu-u^*+k}\binom{n_1}{u^*-2k}\binom{n_2}{k}}{\binom{n}{n\mu}},$$

where the sum is taken over all triples (v_0, v_1, v_2) of integers in the set S defined by $0 \le v_j \le n_j$ for $j = 0, 1, 2$, $v_0 + v_1 + v_2 = n\mu$ and $v_1 + 2v_2 = u^*$, and $u^* = u + (n_1 + 2n_2)\mu$.

Observation 3. Consider a logistic regression model as in (1), where logit $\mu_i = \beta_0 + \beta_1 x_i + \gamma g_i$, and x_i is a binary covariate taking value 0 or 1. Let l_j be the number of individuals with $x_i = 0$ and genotype $g_i = j,\ j = 0, 1, 2$, and let $l = l_0 + l_1 + l_2$. Define similar counts m_j and m for individuals with $x_i = 1$. Let logit $\mu_0 = \beta_0$, and logit $\mu_1 = \beta_0 + \beta_1$. Then, under the null hypothesis,

$$P(U_\gamma = u \mid \boldsymbol{U}_\beta = \boldsymbol{0}) = \sum_{s \in S} \frac{\binom{l_0}{v_0}\binom{l_1}{v_1}\binom{l_2}{v_2}}{\binom{l}{l\mu_0}} \frac{\binom{m_0}{w_0}\binom{m_1}{w_1}\binom{m_2}{w_2}}{\binom{m}{m\mu_1}},$$

where the sum is taken over all sextuples $\boldsymbol{s} = (v_0, v_1, v_2, w_0, w_1, w_2)$ of integers in the set S defined by $0 \le v_j \le l_j,\ 0 \le w_j \le m_j$ for $j = 0, 1, 2,\ v_0 + v_1 + v_1 = l\mu_0,\ w_0 + w_1 + w_2 = m\mu_1$ and $v_1 + 2v_2 - (l_1 + 2l_2)\mu_0 + w_1 + 2w_2 - (m_1 + 2m_2)\mu_1 = u$.

The exact conditional score test defined in Observation 2 reduces to Fisher's exact test for 2×2 contingency tables when $n_2 = 0$. From Observations 2 and 3, it follows that an exact p-value can be computed for special cases of the logistic regression model (1). An extension of Observation 3 can also be derived for regression models with several categorical nuisance covariates. However, for more complex covariate patterns, this approach becomes computationally infeasible, and even intractable when continuous covariates are included. Section 3 introduces methods for computing p-values under conditional inference using saddlepoint approximations.

2.2.3 | Two-sided p-values and mid-p-values

In the following, we obtain two-sided p-values by computing exactly or estimating $P(|U_\gamma| \ge |u| \mid \boldsymbol{U}_\beta = \boldsymbol{0})$ (under the null) for some observation u. For an observation $u > 0$, we define the opposite grid point on the lattice defined in Observation 1, u_{opp}, as the nearest grid point to $-u$ from the left. If u_{opp} lies within the support of the conditional distribution (Appendix A), two-sided p-values are defined by $P(U_\gamma \ge u \mid \boldsymbol{U}_\beta = \boldsymbol{0}) + P(U_\gamma \le u_{\text{opp}} \mid \boldsymbol{U}_\beta = \boldsymbol{0})$, and otherwise $P(U_\gamma \ge u \mid \boldsymbol{U}_\beta = \boldsymbol{0})$. Two-sided mid-$p$ values are defined by $\frac{1}{2}P(U_\gamma \ge u \mid \boldsymbol{U}_\beta = \boldsymbol{0}) + \frac{1}{2}P(U_\gamma \ge u+1 \mid \boldsymbol{U}_\beta = \boldsymbol{0}) + \frac{1}{2}P(U_\gamma \le u_{\text{opp}} \mid \boldsymbol{U}_\beta = \boldsymbol{0}) + \frac{1}{2}P(U_\gamma \le u_{\text{opp}} - 1 \mid \boldsymbol{U}_\beta = \boldsymbol{0})$, or $\frac{1}{2}P(U_\gamma \ge u \mid \boldsymbol{U}_\beta = \boldsymbol{0}) + \frac{1}{2}P(U_\gamma \ge u + 1 \mid \boldsymbol{U}_\beta = \boldsymbol{0})$ if u_{opp} is outside the grid. A similar procedure holds for negative observations $u < 0$, then defining u_{opp} as the nearest grid point to $-u$ from the right.

3 | SADDLEPOINT APPROXIMATIONS

Saddlepoint approximations to cumulative distribution functions are often accurate in situations when the normal approximation might have been used, and - relevant for our situation - can be accurate far into the tails of a distribution [13]. Below, we consider saddlepoint methods for approximating the cumulative distribution of U_γ given $\boldsymbol{U}_\beta = \boldsymbol{0}$ as well as the univariate distribution of the efficient score $\tilde{U}_\gamma$.

3.1 | Double saddlepoint approximation

Conditional tail probabilities $P(U_\gamma \ge u \mid \boldsymbol{U}_\beta = \boldsymbol{0})$ may be estimated by *double saddlepoint approximation* [13]. This will require the *cumulant generating function* of $\boldsymbol{U}$ and $\boldsymbol{U}_\beta$ from Equation (3). The joint cumulant generating function of $\boldsymbol{U}$ is defined by $K(\boldsymbol{t}) = \ln E\left(e^{\boldsymbol{t}^{\mathrm{T}}\boldsymbol{U}}\right)$, were $\boldsymbol{t}$ is a vector of dimension $d + 1$. By using the fact that Y_i is Bernoulli distributed with parameter μ_i, we obtain

$$K(\boldsymbol{t}) = \sum_{i=1}^{n} \left(\ln\left(1 - \mu_i + \mu_i e^{\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}_i}\right) - \mu_i \boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}_i \right), \tag{7}$$

$$\nabla K(\boldsymbol{t}) = \sum_{i=1}^{n} \mu_i \left(\frac{1}{(1-\mu_i)e^{-\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}_i} + \mu_i} - 1 \right) \boldsymbol{z}_i, \quad \text{and} \tag{8}$$

$$H(\boldsymbol{t}) = \sum_{i=1}^{n} \frac{\mu_i(1-\mu_i)e^{-\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}_i}}{\left((1-\mu_i)e^{-\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}_i} + \mu_i\right)^2} \boldsymbol{z}_i \boldsymbol{z}_i^{\mathrm{T}}, \tag{9}$$

where ∇K and H denote the gradient and the Hessian of K, respectively, and $\boldsymbol{z}_i = \left(\boldsymbol{x}_i^{\mathrm{T}}\ g_i\right)^{\mathrm{T}}$. The cumulant generating function of $\boldsymbol{U}_\beta$, its gradient and Hessian, K_β, ∇K_β and H_β, respectively, are obtained by replacing $\boldsymbol{z}_i$ by $\boldsymbol{x}_i$ and letting $\boldsymbol{t}$ have dimension d in (7)–(9). The survival function (right-tail probability) $S(u) = P(U_\gamma \ge u \mid \boldsymbol{U}_\beta = \boldsymbol{0})$ can be approximated as given by

Barndorff-Nielsen[14],

$$\hat{S}(u) = 1 - \Phi\Big(w - \frac{1}{w}\ln\frac{v}{w}\Big), \tag{10}$$

where Φ denotes the standard normal cumulative distribution function. Since U_γ is lattice random variable we use the so-called second continuity correction of Daniels[15] to define w and v. Using $f(\boldsymbol{t}_1, \boldsymbol{t}_2)$ as shorthand for $f\left(\left(\boldsymbol{t}_1^{\mathrm{T}}\ \boldsymbol{t}_2^{\mathrm{T}}\right)^{\mathrm{T}}\right)$, where f is a function and $\boldsymbol{t}_1$, $\boldsymbol{t}_2$ vectors, we have

$$\begin{aligned} w &= \mathrm{sgn}(\hat{t}_\gamma)\sqrt{2\Big(-K(\hat{\boldsymbol{t}}_\beta, \hat{t}_\gamma) + \hat{t}_\gamma\Big(u - \frac{1}{2}\Big)\Big)} \quad \text{and} \\ v &= 2\Big(\sinh\frac{\hat{t}_\gamma}{2}\Big)\sqrt{\frac{\det H(\hat{\boldsymbol{t}}_\beta, \hat{t}_\gamma)}{\det H_\beta(\mathbf{0})}}, \end{aligned} \tag{11}$$

where $\left(\hat{\boldsymbol{t}}_\beta^{\mathrm{T}}\ \hat{t}_\gamma\right)^{\mathrm{T}}$ is the *saddlepoint* satisfying $\nabla K(\hat{\boldsymbol{t}}_\beta, \hat{t}_\gamma) = \left(\mathbf{0}^{\mathrm{T}}\ u - 1/2\right)^{\mathrm{T}}$(Skovgaard[16], see Butler[13], p.114). In general, also the d-dimensional vector $\tilde{\boldsymbol{t}}_\beta$ satisfying $\nabla K_\beta(\tilde{\boldsymbol{t}}_\beta) = \mathbf{0}$ is involved in the expressions for w and v, but $\tilde{\boldsymbol{t}}_\beta = \mathbf{0}$ in our case (see Appendix C). Left-tail probabilities can be approximated, taking into account that U_γ is a lattice variable with step 1, by $1 - \hat{S}(u+1)$.

3.1.1 | Two-sided p-values and mid-p-values

As previously mentioned, for an observation $u > 0$, we define the opposite grid point u_{opp} as the nearest grid point to $-u$ from the left. Assuming u_{opp} lies within the bounds of the lattice distribution, we compute a two-sided p-value by $\hat{S}(u) + 1 - \hat{S}(u_{\mathrm{opp}} + 1)$, where the first term involves the saddlepoint $\left(\hat{\boldsymbol{t}}_\beta^{\mathrm{T}}\ \hat{t}_\gamma\right)^{\mathrm{T}}$ satisfying $\nabla K(\hat{\boldsymbol{t}}_\beta, \hat{t}_\gamma) = \left(\mathbf{0}^{\mathrm{T}}\ u - 1/2\right)^T$ and the latter term involves the saddlepoint $\left(\hat{\boldsymbol{t}}_\beta^{\mathrm{T}}\ \hat{t}_\gamma\right)^{\mathrm{T}}$ satisfying $\nabla K(\hat{\boldsymbol{t}}_\beta, \hat{t}_\gamma) = \left(\mathbf{0}^{\mathrm{T}}\ u_{\mathrm{opp}} + 1/2\right)^T$. For negative observations $u < 0$, the two-sided p-value is given by $\hat{S}(u_{\mathrm{opp}}) + 1 - \hat{S}(u+1)$. Mid-$p$-values may be obtained using double saddlepoint methods *without* continuity corrections (see e.g. Butler[13], p.188). We compute mid-p-values by $\hat{S}(u) + 1 - \hat{S}(u_{\mathrm{opp}})$, but now using

$$w = \mathrm{sgn}(\hat{t}_\gamma)\sqrt{2\Big(-K(\hat{\boldsymbol{t}}_\beta, \hat{t}_\gamma) + \hat{t}_\gamma u\Big)} \quad \text{and} \quad v = \hat{t}_\gamma\sqrt{\frac{\det H(\hat{\boldsymbol{t}}_\beta, \hat{t}_\gamma)}{\det H_\beta(\mathbf{0})}},$$

instead of (11), and where $\left(\hat{\boldsymbol{t}}_\beta^{\mathrm{T}}\ \hat{t}_\gamma\right)^{\mathrm{T}}$ is the saddlepoint satisfying $\nabla K(\hat{\boldsymbol{t}}_\beta, \hat{t}_\gamma) = \left(\mathbf{0}^{\mathrm{T}}\ u\right)^{\mathrm{T}}$, or $\nabla K(\hat{\boldsymbol{t}}_\beta, \hat{t}_\gamma) = \left(\mathbf{0}^{\mathrm{T}}\ u_{\mathrm{opp}}\right)^{\mathrm{T}}$.

3.2 | Single saddlepoint approximation using the efficient score

The above method is related to the test proposed by Dey et al.[4], which may be interpreted as a saddlepoint approximation to the distribution of the efficient score (Equation 6). The efficient score can be expressed as $\tilde{U}_\gamma = \tilde{\boldsymbol{g}}^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{\mu})$ with $\tilde{\boldsymbol{g}} = \boldsymbol{g} - X(X^T W X)^{-1} X^T W \boldsymbol{g}$ the vector of all components $\tilde{g}_i$, and defined as in Dey et al[4]. Using the normal assumption, the unconditional distribution of the efficient score with $\hat{\boldsymbol{\mu}}$ treated as a plug-in constant is the same as the conditional distribution of $U_\gamma | \boldsymbol{U}_\beta = \mathbf{0}$. Therefore, it is of interest to study whether inference based on the univariate distribution of $\tilde{U}_\gamma(\hat{\boldsymbol{\mu}})$ as estimated by a single saddlepoint approach and with $\hat{\boldsymbol{\mu}}$ treated as a plug-in constant, resembles conditional inference of U_γ given $\boldsymbol{U}_\beta = \mathbf{0}$ based on the double saddlepoint approach described above. Tail probabilities of the efficient score may be approximated by a single saddlepoint method via the univariate cumulant generating function of $\tilde{U}_\gamma$, given by

$$K(t) = \sum_{i=1}^{n} \ln(1 - \hat{\mu}_i + \hat{\mu}_i e^{\tilde{g}_i t}) - t\,\tilde{\boldsymbol{g}}^T\hat{\boldsymbol{\mu}},$$

and similarly to the continuity-corrected double saddlepoint method outlined in the previous section, we propose that left-tail probabilities should be estimated as in Equation (10), now with

$$w = \mathrm{sgn}(\hat{t})\sqrt{2(\hat{t}(u - 1/2) - K(\hat{t}))}, \text{ and } v = 2\sinh(\hat{t}/2)\sqrt{K''(\hat{t})},$$

where $\hat{t}$ is the saddlepoint obtained by solving $K'(\hat{t}) = u - 1/2$. Since such a two-step approach does not require a double saddlepoint approximation, this method will require less computational time. Methods for computing two-sided p-values and mid-p-values follow from the previous discussion.

4 | COMPARISON OF METHODS

4.1 | Intercept model

Our aim is to develop approximation methods to compute p-values in situations where the normal assumption is inaccurate and exact tests are intractable or unavailable. However, we find it useful to first study the properties of such approximation methods in situations where an exact test is available, namely for the intercept model. For a specified significance level α, a valid test satisfies $P(\text{type I error}) \leq \alpha$. Under the null, $Y_i \sim \text{binom}(1, \mu)$ for all $i = 1, \dots, n$, where $\mu = \exp(\beta_0)/(1 + \exp(\beta_0))$. For a particular realization $\boldsymbol{y}$, and using a conditional score test, we compare the observed score test statistic $u = \boldsymbol{g}^T(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) = \boldsymbol{g}^T(\boldsymbol{y} - \bar{y}\mathbf{1})$ to the null distribution of U_γ given $\sum_i Y_i = n\bar{y} = v$. For all datasets in which the response vector $\boldsymbol{y}$ satisfies $\sum_i y_i = v$, tests are *conditionally* valid when $P(\text{type I error}| \sum_i Y_i = v) \leq \alpha$. The *overall* probability of type I error for a specified null model (fixed μ and $\boldsymbol{g}$) is given by

$$P(\text{type I error}) = \sum_v \left[P\left(\text{type I error} \mid \sum_{i=1}^n Y_i = v \right) P\left(\sum_{i=1}^n Y_i = v \right) \right]. \tag{12}$$

As the exact test is conditionally valid for all v, it will also be valid overall. Approximation methods may be conditionally valid for some v, but invalid overall, or valid overall but conditionally invalid for some v.

We will illustrate the methods presented in Sections 3 and 4 with a simple numerical example. Let $n = 10\,000$, define $\boldsymbol{g}$ such that $n_0 = 8100$, $n_1 = 1800$ and $n_2 = 100$ (MAF $= 0.1$) and consider significance level $\alpha = 5 \cdot 10^{-8}$. For each $v \in \{1, \dots n-1\}$, we use the exact conditional distribution of Observation 2 to obtain the largest critical region so that the exact two-sided test is conditionally valid, and in a similar fashion we use approximation methods to establish critical regions. Conditional type I error rates of approximation methods are calculated using the exact conditional distribution on these critical regions. For a given μ, the overall type I error rate may then be calculated as in Equation (12).

For $\mu \in (0, 0.5)$, overall type I error rates of the exact test and the saddlepoint approximations described above (double saddlepoint with continuity correction, DSPA-CC, and single saddlepoint approximation based on the efficient score with continuity correction, ESPA-CC) are plotted in Figure 1. Type I error rates of the normal approximation are also plotted. We first observe that the exact test is always conservative, an observation which of course is well-known for discrete test statistics. Further, we observe that the DSPA-CC method closely resembles the exact test, while the ESPA-CC method is somewhat more conservative for $\mu < 0.2$ in this example. The normal approximation is clearly inaccurate and anti-conservative for $\mu < 0.3$.

Saddlepoint approximation methods *without* continuity corrections will give less (approximately $1/2$) weight to the observation u (and the opposite grid point) when calculating a p-value, and can therefore be considered approximations of exact mid-p-values in this setting. Type I error rates of such methods (denoted DSPA and ESPA) and exact mid-p-values are also presented in Figure 1. The DSPA method closely resembles the exact test with mid-p, although the type I error rate is slightly higher for small values of μ. The ESPA method is somewhat more conservative than the exact mid-p method. In this example, the mid-p method is at times slightly anti-conservative, and so are the approximation methods. Although the *overall* type I error rates lie close to the desired significance level $\alpha = 5 \cdot 10^{-8}$, Figure 1 illustrates that these methods will often be *conditionally* invalid.

4.2 | Simulations resembling genetic association studies with an imbalanced response

The purpose of the following simulation study is to compare methods in a setting resembling a genome-wide association study with an imbalanced response, for which exact tests are not available. The simulation set-up is motivated by Dey et al.[4] and we estimate type I error rates conditional on the number of cases. The sample size considered is $n = 20000$ of which 2% are cases. We consider the logistic regression model

$$\text{logit}(\mu_i) = \beta_0 + x_{1i} + x_{2i} + \gamma g_i,$$

with $X_1 \sim \text{Bernoulli}(0.5)$, $X_2 \sim N(0, 1)$ and $G \sim \text{binom}(2, m)$ with the parameter m (representing a minor allele frequency) taking the values 0.05, 0.005, 0.0005 and 0.00025. Since we are evaluating validity of tests we set $\gamma = 0$, and we set $\beta_0 = -5.6$ so that the disease prevalence is 1% in the population. The covariates x_{1i} and x_{2i} are sampled conditionally on the phenotype value y_i, while the genotype value is sampled independently of this under the null hypothesis. See Supplementary File for details. Holding $\boldsymbol{y}$ fixed and sampling the covariates ensures that the number of cases is equal for all simulations. For each minor allele frequency (0.05, 0.005, 0.0005 or 0.00025), we simulate 10^9 data sets and record the number of times the null hypothesis is rejected at the $\alpha = 5 \cdot 10^{-8}$ significance level when using the double saddlepoint approximation with continuity correction

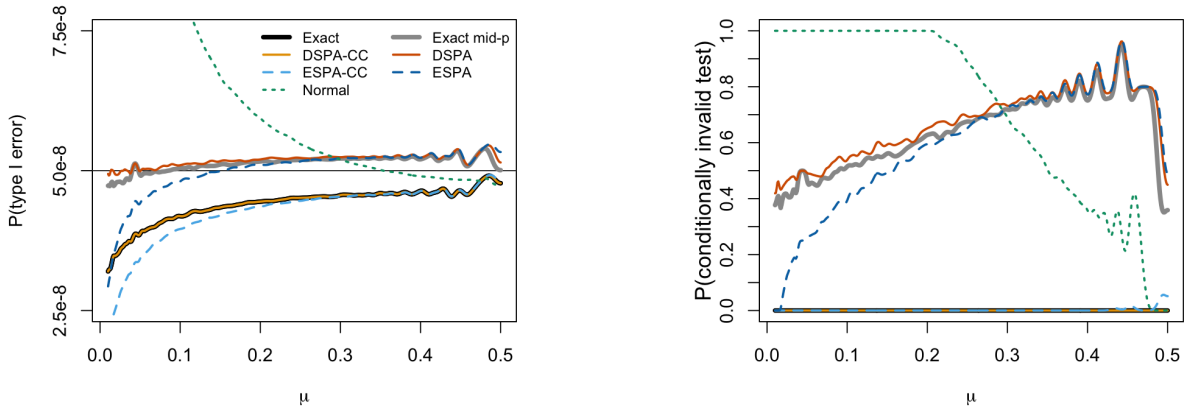

FIGURE 1 Left: Overall type I error rates for the intercept model where $n = 10\,000$, $n_0 = 8100$, $n_1 = 1800$ and $n_2 = 100$. Two-sided critical regions on the lattice of $U_\gamma | U_{\beta_0} = 0$ are determined (for the exact test) or approximated (using saddlepoint methods and the normal approximation) based on the criterion $P(\text{type I error}) \leq 5 \cdot 10^{-8}$. Right: The probability of sampling a response ($\boldsymbol{Y}$) under the null such that the conditional type I error rate exceeds the significance level $\alpha = 5 \cdot 10^{-8}$, or equivalently the proportion of conditionally invalid tests for each value of μ.

(DSPA-CC), the continuity-corrected single saddlepoint approximation based on the efficient score (ESPA-CC), as well as the double and single saddlepoint approximations without continuity corrections (DSPA and ESPA). The resulting empirical type I error rates are presented in Figure 2, along with 95% Clopper-Pearson confidence intervals. The simulation results resemble some of the observations made in the previous section. The DSPA-CC and ESPA-CC are always conservative, while the mid-p methods (DSPA and ESPA) are less conservative, or even anti-conservative for small minor allele frequencies. Methods based on the efficient score are somewhat more conservative than those based on the double saddlepoint method. An additional simulation example is presented in the Supplementary File.

5 | APPLICATION TO UK BIOBANK DATA

5.1 | Common variants

We consider a recent GWAS in the UK Biobank with motivation from Rogne et al.[5] The phenotype of interest is skin and soft tissue infections (SSTIs), and individuals are defined as cases if they have been hospitalized with main ICD-10 codes A46 (erysipelas), L03 (cellulitis and acute lymphangitis), or M72.6 (necrotizing fasciitis) in the period between the end of the recruitment period (2010-10-01) and April 2017 (2017-03-31). Individuals who had reported ICD-10 codes, or corresponding ICD-9 codes (035 and 729.4), before 2010-10-01 are removed as well as individuals with date of death reported after 2010-10-01 in the death register (see Data-Field 40000 in the UK Biobank data). As nuisance covariates we include age when attended assessment centre, genetic sex, and four principal components. To avoid complexities due to cryptic relatedness we only include unrelated individuals reported as White British (achieved through Data-Field 22006 and 22020 in UK Biobank). The principal components are calculated using EIGENSOFT (version 6.1.4) SmartPCA.[17,18] Only directly genotyped SNPs are considered, and phenotype-independent quality control of the genetic data is completed using PLINK1.9, with details given in the Supplementary File. This results in a total of 293 964 individuals and 529 024 SNPs with 2051 individuals defined as cases and 291 913 controls, resulting in a case proportion of 0.7 %.

All SNPs were first investigated by computing p-values using the normal approximation to the score test statistic. As this test is proven to be too optimistic, SNPs with p-values less than $\alpha = 5 \cdot 10^{-5}$ were investigated more thoroughly by using the DSPA-CC and ESPA-CC methods as implemented by us, as well as the SPA-test of Dey et al.[4]. This test should resemble the ESPA method used in simulations in the previous section. In Dey et al.[4], a computationally more efficient approximation to their SPA-test is also proposed by essentially assuming that the nuisance covariates are balanced. In a double saddlepoint setting, this

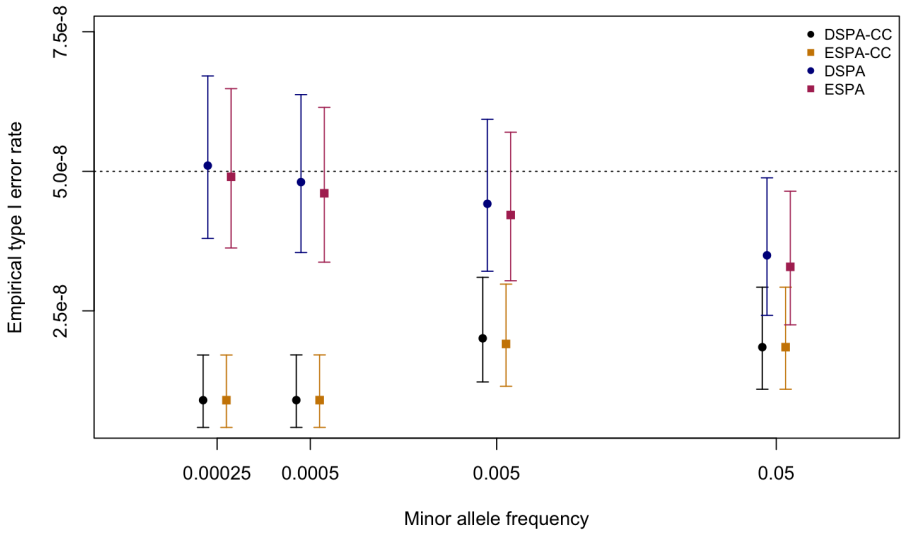

FIGURE 2 Empirical type I error rates for simulations with covariates using the double saddlepoint method with continuity correction (DSPA-CC), the single saddlepoint method based on the efficient score with continuity correction (ESPA-CC) and, for comparison, the double saddlepoint method without continuity correction (DSPA) and the single saddlepoint method based on the efficient score without continuity correction (ESPA). Error rates are estimated for genetic variants with minor allele frequencies 0.00025, 0.0005, 0.005, and 0.05.

assumption may be generalized to argue that the score vector $\boldsymbol{U}_{\boldsymbol{\beta}}$ approximately has a multivariate normal distribution under the null hypothesis. By taking a similar approach, we may partition the joint CDF of $\boldsymbol{U}_{\boldsymbol{\beta}}$ and U_γ into a sum over all individuals with genotype value $g_i > 0$ and those with $g_i = 0$. For the latter sub-sample, the CGF simplifies to a CGF of the score vector $\boldsymbol{U}^*_{\boldsymbol{\beta}}$ including individuals with $g_i = 0$. Assuming that also $\boldsymbol{U}^*_{\boldsymbol{\beta}}$ is normal, this part of the joint CGF may be replaced by a normal CGF, and by pre-computing the variance of $\boldsymbol{U}^*_{\boldsymbol{\beta}}$, an approximated double saddlepoint method may be computed based only on the sub-sample of individuals with genotypes $g_i > 0$. Details may be found in the Supplementary File. For comparative purposes, we compute p-values based on the fastSPA method of Dey et al.[4] and our related fastDSPA-CC approach.

Test results for the SNPs with the smallest normal-approximated p-values are given in Table 1. In this setting, we no longer know whether the null hypothesis is true or not for each variant. However, we expect only a tiny proportion of all variants where the null hypothesis is false. Even though no SNPs reached the significance level $\alpha = 5 \cdot 10^{-8}$, we see a pattern similar to our previous simulation results. The normal approximation is the most optimistic, followed by the SPA and fastSPA tests which give mid-p-values. The DSPA-CC test is more conservative, while the most conservative test is ESPA-CC. The fastDSPA-CC method is slightly less conservative than DSPA-CC. As would be expected based on the aforementioned simulation results, the greatest difference between tests is observed for the SNP with a low minor allele frequency (rs113113104, MAF = 0.03). The difference between the p-values reduces for increasing MAFs. For the SNP rs566530 with MAF = 0.48, the SPA test gives a smaller p-value than the normal approximation, while the other methods give larger p-values.

5.2 | Rare variants

We have so far observed that the difference between approximation methods is most pronounced when minor allele frequencies are low, and the response is highly skewed. To further study such situations, we consider the UK Biobank exome sequence data consisting of 45 596 unrelated individuals of European origin. We limit ourselves to White British individuals using the same requirements for the definition of SSTIs as for the common variants. This results in a total number of 30 210 individuals to investigate with 210 individuals defined as cases, once again leading to a case proportion of about 0.7 %. See the Supplementary File for further information about quality control. The principal components are computed as for the common variants analysis,

TABLE 1 The common variants with the smallest computed p-values using the normal approximation to the score test statistic for the GWAS of skin and soft tissue infections. Results from saddlepoint approximation methods are included for comparison.

SNP	CHR	MAF	Norm	SPA	fastSPA	ESPA-CC	DSPA-CC	fastDSPA-CC
rs113113104	6	0.03	$2.39{\cdot}10^{-7}$	$5.97{\cdot}10^{-7}$	$6.04{\cdot}10^{-7}$	$7.27{\cdot}10^{-7}$	$7.10{\cdot}10^{-7}$	$6.52{\cdot}10^{-7}$
rs6551253	3	0.28	$8.38{\cdot}10^{-6}$	$8.47{\cdot}10^{-6}$	$8.78{\cdot}10^{-6}$	$9.18{\cdot}10^{-6}$	$9.00{\cdot}10^{-6}$	$8.92{\cdot}10^{-6}$
rs78404737	2	0.10	$8.50{\cdot}10^{-6}$	$9.63{\cdot}10^{-6}$	$9.78{\cdot}10^{-6}$	$1.08{\cdot}10^{-5}$	$1.06{\cdot}10^{-5}$	$1.00{\cdot}10^{-5}$
rs78696065	7	0.02	$8.80{\cdot}10^{-6}$	$1.54{\cdot}10^{-5}$	$1.55{\cdot}10^{-5}$	$1.89{\cdot}10^{-5}$	$1.87{\cdot}10^{-5}$	$1.75{\cdot}10^{-5}$
rs479947	6	0.11	$1.19{\cdot}10^{-5}$	$1.29{\cdot}10^{-5}$	$1.33{\cdot}10^{-5}$	$1.44{\cdot}10^{-5}$	$1.42{\cdot}10^{-5}$	$1.35{\cdot}10^{-5}$
rs566530	6	0.48	$1.46{\cdot}10^{-5}$	$1.40{\cdot}10^{-5}$	$1.48{\cdot}10^{-5}$	$1.50{\cdot}10^{-5}$	$1.47{\cdot}10^{-5}$	$1.48{\cdot}10^{-5}$
rs56355912	10	0.03	$1.51{\cdot}10^{-5}$	$2.16{\cdot}10^{-5}$	$2.16{\cdot}10^{-5}$	$2.57{\cdot}10^{-5}$	$2.54{\cdot}10^{-5}$	$2.38{\cdot}10^{-5}$
rs72733294	5	0.36	$1.58{\cdot}10^{-5}$	$1.60{\cdot}10^{-5}$	$1.60{\cdot}10^{-5}$	$1.72{\cdot}10^{-5}$	$1.69{\cdot}10^{-5}$	$1.69{\cdot}10^{-5}$
rs11074743	16	0.40	$1.69{\cdot}10^{-5}$	$1.68{\cdot}10^{-5}$	$1.71{\cdot}10^{-5}$	$1.80{\cdot}10^{-5}$	$1.77{\cdot}10^{-5}$	$1.77{\cdot}10^{-5}$
rs1562963	11	0.07	$2.02{\cdot}10^{-5}$	$1.99{\cdot}10^{-5}$	$2.33{\cdot}10^{-5}$	$2.26{\cdot}10^{-5}$	$2.23{\cdot}10^{-5}$	$2.13{\cdot}10^{-5}$

however separately on these 30 210 individuals. We have only considered rare variants in chromosome 6 with a minor allele count (MAC) of 3 or higher. The results of single variant tests applied to rare variants are given in Table 2.

TABLE 2 The rare variants with the smallest computed p-values using the normal approximation to the score test statistic for the GWAS of skin and soft tissue infections. Results from saddlepoint approximation methods are included for comparison.

SNP	CHR	MAC	Norm	SPA	fastSPA	ESPA-CC	DSPA-CC	fastDSPA-CC
6:26045407:G:A	6	4	$2.07{\cdot}10^{-36}$	$4.31{\cdot}10^{-5}$	$4.31{\cdot}10^{-5}$	$2.20{\cdot}10^{-4}$	$2.20{\cdot}10^{-4}$	$2.20{\cdot}10^{-4}$
6:41097421:T:C	6	4	$2.21{\cdot}10^{-32}$	$4.92{\cdot}10^{-5}$	$4.92{\cdot}10^{-5}$	$2.60{\cdot}10^{-4}$	$2.60{\cdot}10^{-4}$	$2.50{\cdot}10^{-4}$
6:24852645:G:T	6	4	$1.37{\cdot}10^{-25}$	$8.93{\cdot}10^{-5}$	$8.93{\cdot}10^{-5}$	$4.40{\cdot}10^{-4}$	$4.30{\cdot}10^{-4}$	$4.20{\cdot}10^{-4}$
6:31772925:C:A	6	5	$6.36{\cdot}10^{-23}$	$1.30{\cdot}10^{-4}$	$1.30{\cdot}10^{-4}$	$6.00{\cdot}10^{-4}$	$6.00{\cdot}10^{-4}$	$5.80{\cdot}10^{-4}$
6:20402579:C:T	6	3	$4.19{\cdot}10^{-22}$	$2.00{\cdot}10^{-3}$	$2.00{\cdot}10^{-3}$	$1.00{\cdot}10^{-2}$	$1.00{\cdot}10^{-2}$	$1.00{\cdot}10^{-2}$
6:132588925:C:T	6	6	$8.78{\cdot}10^{-22}$	$1.50{\cdot}10^{-4}$	$1.50{\cdot}10^{-4}$	$6.90{\cdot}10^{-4}$	$6.90{\cdot}10^{-4}$	$6.70{\cdot}10^{-4}$
6:17675831:G:A	6	3	$8.94{\cdot}10^{-22}$	$2.20{\cdot}10^{-3}$	$2.20{\cdot}10^{-3}$	$1.00{\cdot}10^{-2}$	$1.00{\cdot}10^{-2}$	$1.00{\cdot}10^{-2}$
6:110960684:T:G	6	3	$2.05{\cdot}10^{-21}$	$1.70{\cdot}10^{-3}$	$1.70{\cdot}10^{-3}$	$4.90{\cdot}10^{-3}$	$4.90{\cdot}10^{-3}$	$4.90{\cdot}10^{-3}$
6:7894854:T:C	6	16	$1.88{\cdot}10^{-20}$	$3.07{\cdot}10^{-5}$	$3.07{\cdot}10^{-5}$	$1.20{\cdot}10^{-4}$	$1.20{\cdot}10^{-4}$	$1.00{\cdot}10^{-4}$
6:148514044:G:T	6	3	$1.94{\cdot}10^{-20}$	$2.20{\cdot}10^{-3}$	$2.20{\cdot}10^{-3}$	$1.10{\cdot}10^{-2}$	$1.10{\cdot}10^{-2}$	$1.10{\cdot}10^{-2}$

Clearly, the normal approximation to the score test statistic can be very inaccurate in this setting. We also observe that the saddlepoint approximations with continuity correction can differ from the SPA test (mid-p) in about one order of magnitude. As a result, we expect the use of the continuity correction to be most consequential for rare variants. Another observation is that ESPA-CC and DSPA-CC are practically identical in this application. The speed-up approximation methods are more accurate for rare variants, which can be explained by the fact that the accuracy of the multivariate normal approximation of $\mathbf{U}^*_{\boldsymbol{\beta}}$ in fastDSPA-CC, depends on the number of individuals with $g_i = 0$, which increases for decreasing MACs.

6 | DISCUSSION

We have investigated different saddlepoint approximations for GWAS with binary phenotypes and in particular showed how to apply the double saddlepoint method to approximate conditional tail probabilities for the score test statistic. We have also illustrated the use of a double saddlepoint approach without continuity correction as a means to compute two-sided mid-p-values. The continuity-corrected double saddlepoint approximation, DSPA-CC, and single-saddlepoint approximation based on the efficient score, ESPA-CC, are both found to perform well and are at times remarkably similar. However there are situations in which ESPA-CC is somewhat more conservative than DSPA-CC and so DSPA-CC will be more powerful.

The methods considered here are not limited to genetic association studies, but may be considered for other applications where both the response variable and the covariate of interest are imbalanced. However, some comments should be made with regards to the use of these methods in genetic association studies. The use of a continuity corrected saddlepoint approach was

motivated by Observation 1 in which the covariate of interest, g, takes values 0, 1 or 2. Due to the discrete nature of the test statistic, these tests are conservative. And if applied genome-wide, discrete test statistics (such as DSPA-CC and ESPA-CC) will result in deflated QQ-plots, which in turns makes the QQ-plot less effective for checking model assumptions. The mid-p methods (DSPA and ESPA), which are saddlepoint approximations without continuity corrections, could be better suited, although neither approach is guaranteed to control the type I error rate conditionally. Furthermore, is not uncommon to use imputed SNPs in GWAS. With most imputation methods, the output for each imputed SNP is a probability that the minor allele count is equal to 0, 1 or 2, denoted p_0, p_1 and p_2. If the imputed genotype is set to be the expected minor allele count, $p_1 + 2p_2$, the score test statistic will no longer have a lattice distribution, and so the proposed continuity correction becomes inaccurate. A work-around could be to set the imputed minor allele count equal to the most likely allele count. Otherwise, the continuous saddlepoint approximations such as DSPA and ESPA (coinciding with the SPA-test of Dey et al[4]) could be more suitable.

One well-known advantage of the asymptotic score test over the likelihood ratio and Wald tests is that the test statistic only includes parameter estimates for the null model. In genome-wide testing, the null model is the same for all tests and so the score test is particularly appealing. However, the saddlepoint equations must be solved numerically for each genetic variant, which can make the procedure computationally slow. The ESPA-CC test is considerably faster to compute, as it relies only on a single saddlepoint method, and the method of Dey et al includes further alternatives for fast computations.[4] When applied to UK Biobank data in Section 6 we did not compute p-values using saddlepoint methods on all SNPs, but rather proposed to use our methods to adjust the top results obtained by standard methods. With such an approach, QQ-plots may be plotted using standard outputs of statistical software for GWAS. However, as the normal approximation would result in inflated QQ-plots in imbalanced data sets, we would rather suggest to use continuous saddlepoint approximations (such as DSPA or SPA-test of Dey et al) for model diagnostics. At the end of Section 6.2 we applied our methods for single-variant tests on rare variants. Such tests are often low-powered and region-based tests are more commonly applied for rare variants. However, many of these methods rely on single-variant tests as building blocks, among them SAIGE-GENE+ and ACAT[19,20] which incorporate single saddlepoint methods based on the efficient score. Using double saddlepoint methods in region-based tests for logistic regression models with an imbalanced response could be a topic for future research.

Some comments should also be made regarding details of the saddlepoint methods. An alternative saddlepoint approximation $\hat{S}(u)$ to the cumulative distribution of a random variable is the one introduced in Lugannani and Rice[21]. This approximation gives the same results as the approximation by Barndorff-Nielsen[14] in many situations, but we observed that the approximation by Lugananni and Rice was inaccurate in simulations when the case proportion and minor allele frequency approached zero. See for instance Booth and Wood[22] for similar observations in a different application. Butler[13] lists three continuity correction methods of Davies[15] for estimating discrete cumulative probabilities using saddlepoint methods, of which the second has been applied by us. Procedures for obtaining two-sided p-values are simpler with this method. The first continuity correction was however also investigated and gave very similar results but was slightly more inaccurate for the intercept model.

7 | ACKNOWLEDGEMENTS

This research was supported by the Norwegian Research Council grant 272402 (PhD Scholarships at SINTEF) as well the funding for research stays abroad for doctoral and postdoctoral fellows financed by the Norwegian Research Council. The research has been conducted using the UK Biobank Resource under Application Number 32285. We thank the Yale Center for Research Computing for guidance and use of the research computing infrastructure. We thank the The Gemini Center for Sepsis Research for establishing cooperation with Yale School of Public Health. We also thank anonymous reviewers for insightful comments.

8 | DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from UK Biobank. Restrictions apply to the availability of these data. Data are available for bona fide researchers through an open application.

9 | CODE AVAILABILITY

Source code in R[23] is available at https://github.com/palVJ/SaddlePointApproxInBinaryGWAS.

□

APPENDIX

A SUPPORT OF THE CONDITIONAL SCORE TEST STATISTIC

Consider the score test statistic of U_γ conditional on $\boldsymbol{U}_\beta = \mathbf{0}$. We have $-\hat{\boldsymbol{\mu}} \le \boldsymbol{Y} - \hat{\boldsymbol{\mu}} \le \mathbf{1} - \hat{\boldsymbol{\mu}}$ (elementwise inequalities), where $\mathbf{1}$ is a vector of ones. Since all $g_i \ge 0$, premultiplying the inequalities with $\boldsymbol{g}^{\mathrm{T}}$ gives bounds on the support of $\boldsymbol{g}^T(\boldsymbol{Y} - \hat{\boldsymbol{\mu}})$:

$$-\boldsymbol{g}^{\mathrm{T}}\hat{\boldsymbol{\mu}} \le U_\gamma \le \boldsymbol{g}^{\mathrm{T}}(\mathbf{1} - \hat{\boldsymbol{\mu}}). \tag{A1}$$

The first equality holds when $\boldsymbol{g}^{\mathrm{T}}\boldsymbol{Y} = 0$ and the second when $\boldsymbol{g}^{\mathrm{T}}\boldsymbol{Y} = \boldsymbol{g}^{\mathrm{T}}\mathbf{1}$. However, this combination is not achievable if it does not satisfy $\boldsymbol{U}_\beta = \boldsymbol{X}^T(\boldsymbol{Y} - \hat{\boldsymbol{\mu}}) = \mathbf{0}$. Specifically, the minimal and maximal achievable values of the conditional score test statistic is given by the constraint optimization problems:

$$\begin{aligned} \min(U_\gamma) = \min_{\boldsymbol{y}} \quad & \boldsymbol{g}^T(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) \\ \text{such that} \quad & X^T(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) = \mathbf{0}, \end{aligned}$$

and

$$\begin{aligned} \max(U_\gamma) = \max_{\boldsymbol{y}} \quad & \boldsymbol{g}^T(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) \\ \text{such that} \quad & X^T(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) = \mathbf{0}. \end{aligned}$$

As an example, consider the intercept model with $n = 1000$ and $\boldsymbol{g}$ such that $n_0 = 980$, $n_1 = 20$ and $n_2 = 0$ and $\boldsymbol{Y}$ such that $\sum_{i=1}^n Y_i = 10$. Then $\hat{\mu}_i = 10/n = 0.01$ satisfies $U_{\beta_0} = \sum_{i=1}^n (Y_i - \mu_i) = 0$ and the minimum achievable value is indeed $\min(U_\gamma) = -\boldsymbol{g}^T\hat{\boldsymbol{\mu}} = -0.2$, since we may have a combination where $Y_i = 0$ for all $g_i > 0$, and still get $\sum_{i=1}^n Y_i = 10$. However, $\max(U_\gamma) = 10 - \boldsymbol{g}^T\hat{\boldsymbol{\mu}} = 9.8$ since $\boldsymbol{g}^T\boldsymbol{Y}$ can be no larger than the combinations where $g_i = 1$ for all $Y_i = 1$, which can only occur ten times in order to satisfy $\sum_{i=1}^n Y_i = 10$.

B PROOFS OF OBSERVATIONS 2 AND 3

Proof of Observation 2. We assume throughout the proof that the null hypothesis is true, $\gamma = 0$. Denote by V_j the sum of responses Y_i among individuals with genotype $g_i = j,\ j = 0, 1, 2$, and let $V = V_0 + V_1 + V_2 = \sum_{i=1}^n Y_i$ be the total sum of responses. With this notation, $U_\gamma = V_1 + 2V_2 - (n_1 + 2n_2)\mu$, and $U_\beta = V - n\mu$, so that the condition $U_\beta = 0$ is equivalent to $V = n\mu$.

The V_j are independent, and V_j is binomially distributed with parameters n_j and $\mu,\ j = 0, 1, 2$, and V is binomially distributed with parameters n and μ. Assume that $v_0 + v_1 + v_2 = n\mu$ with v_j in the support of V_j. Then

$$\begin{aligned} P\big(V_0 = v_0, V_1 = v_1, V_2 = v_2 \mid V = n\mu\big) &= \frac{P(V_0 = v_0)P(V_1 = v_1)P(V_2 = v_2)}{P(V = n\mu)} \\ &= \frac{\binom{n_0}{v_0}\mu^{v_0}(1-\mu)^{n_0 - v_0}\binom{n_1}{v_1}\mu^{v_1}(1-\mu)^{n_1 - v_1}\binom{n_2}{v_2}\mu^{v_2}(1-\mu)^{n_2 - v_2}}{\binom{n}{n\mu}\mu^{n\mu}(1-\mu)^{n - n\mu}} = \frac{\binom{n_0}{v_0}\binom{n_1}{v_1}\binom{n_2}{v_2}}{\binom{n}{n\mu}}, \end{aligned}$$

a trivariate hypergeometric probability.

Now, $P(U_\gamma = u \mid U_\beta = 0) = P\big(V_1 + 2V_2 = u^* \,\big|\, V = n\mu\big)$ can be found by summing the above probabilities over $(v_0, v_1, v_2) \in S$. This gives the first sum of the Observation. The more explicit second version of the sum is obtained by solving the two equations in the definition of S for v_0 and v_1 in terms of $k = v_2$. The limits of the sum is determined by the inequalities in the definition of S. □

Proof of Observation 3. We assume throughout the proof that the null hypothesis is true, $\gamma = 0$. Denote by V_j the sum of responses Y_i among individuals with $x_i = 0$ and genotype $g_i = j,\ j = 0, 1, 2$, and let $V = V_0 + V_1 + V_2$. Define similar sums

W_j and W for individuals with $x_i = 1$. With this notation, $U_\gamma = V_1 + 2V_2 - (l_1 + 2l_2)\mu_0 + W_1 + 2W_2 - (m_1 + 2m_2)\mu_1$, and $\boldsymbol{U}_\beta^{\mathrm{T}} = \big(V + W - l\mu_0 - m\mu_1 \;\; W - m\mu_1\big)$, so that the condition $\boldsymbol{U}_\beta = \mathbf{0}$ is equivalent to $V = l\mu_0$ and $W = m\mu_1$.

All the V_j and W_j are independent, and V_j is binomially distributed with parameters l_j and μ_0, and W_j with parameters m_j and μ_1, $j = 0, 1, 2$. As in the proof of Observation 2, the conditional point probabilites of (V_0, V_1, V_2) given $V = l\mu_0$ and (W_0, W_1, W_2) given $W = m\mu_1$ are trivariate hypergeometric probabilities, and by independence of the two triples, the conditional joint point probability is the product of the two. Then $P(U_\gamma = u \mid \boldsymbol{U}_\beta = \mathbf{0})$ can be found by summing those probabilities over $\boldsymbol{s} \in S$. □

C SOLUTION TO $\nabla_{t_\beta} K_\beta(\tilde{t}_\beta) = \mathbf{0}$

Consider the marginal cumulant generating function of $\boldsymbol{U}_\beta$, defined by $K_\beta(\boldsymbol{t}_\beta)$ (a function of d variables) with

$$K_\beta(\boldsymbol{t}_\beta) = \sum_{i=1}^{n} \ln(1 - \mu_i + \mu_i \exp(\mathbf{x}_i^T \boldsymbol{t}_\beta)) - \boldsymbol{t}_\beta^T X^T \boldsymbol{\mu}, \tag{C2}$$

and corresponding gradient

$$\nabla_{t_\beta} K_\beta(\boldsymbol{t}_\beta) = \sum_{i=1}^{n} \mu_i \mathbf{x}_i \left(\frac{1}{(1 - \mu_i)\exp(-\mathbf{x}_i^T \boldsymbol{t}_\beta) + \mu_i} - 1 \right). \tag{C3}$$

First, one can easily observe that $\tilde{\boldsymbol{t}}_\beta = \mathbf{0}$ is a solution to $\nabla_{t_\beta} K_\beta(\boldsymbol{t}_\beta) = \mathbf{0}$. Second, if one can prove that the CGF is a convex function, then $\tilde{\boldsymbol{t}}_\beta = \mathbf{0}$ is a unique solution to $\nabla_{t_\beta} K_\beta(\boldsymbol{t}_\beta) = \mathbf{0}$.

Proof. Convexity of a cumulant generating function with *any* random variable $\boldsymbol{U}$, $K(\boldsymbol{t}) = \ln E(e^{\boldsymbol{t}^T \boldsymbol{U}})$, in general follows from the Hölder inequality, $E(|X|^c |Y|^{1-c}) \leq (E|X|)^c (E|Y|)^{1-c}$ for all c in $(0, 1)$, where X and Y are random variables. A function f is convex if $f(c\boldsymbol{t}_1 + (1 - c)\boldsymbol{t}_2) \leq c f(\boldsymbol{t}_1) + (1 - c) f(\boldsymbol{t}_2)$ for all c in $(0, 1)$. Now,

$$\begin{aligned} K(c\boldsymbol{t}_1 + (1 - c)\boldsymbol{t}_2) &= \ln E e^{(c\boldsymbol{t}_1 + (1-c)\boldsymbol{t}_2)^{\mathrm{T}} \boldsymbol{U}} = \ln E\left(e^{c\boldsymbol{t}_1^{\mathrm{T}} \boldsymbol{U}} e^{(1-c)\boldsymbol{t}_2^{\mathrm{T}} \boldsymbol{U}}\right) \\ &\leq \ln\left(\left(E e^{\boldsymbol{t}_1^{\mathrm{T}} \boldsymbol{U}}\right)^c \left(E e^{\boldsymbol{t}_2^{\mathrm{T}} \boldsymbol{U}}\right)^{1-c}\right) = c \ln E e^{\boldsymbol{t}_1^{\mathrm{T}} \boldsymbol{U}} + (1 - c) \ln E e^{\boldsymbol{t}_2^{\mathrm{T}} \boldsymbol{U}} \\ &= c K(\boldsymbol{t}_1) + (1 - c) K(\boldsymbol{t}_2), \end{aligned}$$

showing that K is convex. □

References

1. Goeman JJ, Solari A. Multiple hypothesis testing in genomics. *Statistics in medicine* 2014; 33(11): 1946–1978.
2. Cui X, Dickhaus T, Ding Y, Hsu JC. *Handbook of multiple comparisons*. CRC Press . 2021.
3. Ma C, Blackwell T, Boehnke M, Scott LJ. Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genetic Epidemiology* 2013; 37(6): 539–550.
4. Dey R, Schmidt EM, Abecasis GR, Lee S. A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. *American Journal of Human Genetics* 2017; 101(1): 37–49.
5. Rogne T, Liyanarachi KV, Rasheed H, et al. GWAS Identifies LINC01184/SLC12A2 as a Risk Locus for Skin and Soft Tissue Infections.. *J Invest Dermatol* 2021.
6. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika* 1993; 80(1): 27–38.
7. Brazzale AR, Davison AC. Accurate parametric inference for small samples. *Statistical Science* 2008; 23(4): 465–484.
8. Bedrick EJ, Hill JR. An empirical assessment of saddlepoint approximations for testing a logistic regression parameter. *Biometrics* 1992: 529–544.

9. Lancaster HO. Significance tests in discrete distributions. *Journal of the American Statistical Association* 1961; 56(294): 223–234.

10. Lin DY. An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics* 2005; 21(6): 781–787.

11. Bickel PJ, Klaassen CA, Ritov Y, Wellner JA. *Efficient and adaptive estimation for semiparametric models*. 4. Johns Hopkins University Press Baltimore . 1993.

12. Lindsey JK. *Parametric statistical inference*. Oxford University Press . 1996.

13. Butler RW. *Saddlepoint Approximations with Applications*. Cambridge Series in Statistical and Probabilistic Mathematics Cambridge University Press . 2007.

14. Barndorff-Nielsen OE. Approximate Interval Probabilities. *Journal of the Royal Statistical Society. Series B (Methodological)* 1990; 52(3): 485–496.

15. Daniels HE. Tail probability approximations. *International Statistical Review/Revue Internationale de Statistique* 1987: 37–48.

16. Skovgaard IM. Saddlepoint expansions for conditional distributions. *Journal of Applied Probability* 1987; 24(4): 875–887.

17. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 2006; 38(8): 904–909.

18. Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. *PLOS Genetics* 2006; 2(12): e190.

19. Liu Y, Chen S, Li Z, Morrison AC, Boerwinkle E, Lin X. ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. *The American Journal of Human Genetics* 2019; 104(3): 410–421.

20. Zhou W, Bi W, Zhao Z, et al. SAIGE-GENE+ improves the efficiency and accuracy of set-based rare variant association tests. *Nature genetics* 2022; 54(10): 1466–1469.

21. Lugannani R, Rice S. Saddle point approximation for the distribution of the sum of independent random variables. *Advances in Applied Probability* 1980; 12(2): 475–490. doi: 10.2307/1426607

22. Booth JG, Wood ATA. An example in which the Lugannani-Rice saddlepoint formula fails. *Statistics & Probability Letters* 1995; 23(1): 53–61.

23. R Core Team . *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; Vienna, Austria: 2021.