Interval estimation of genetic susceptibility for retrospective case-control studies

BMC Genetics 2004 5:9

Dmitri V. Zaykin[1] (dmitri.v.zaykin@gsk.com), Zhaoling Meng[1,2] (zhaolingm@yahoo.com), Sujit K. Ghosh[2] (sghosh@stat.ncsu.edu)

[1] Department of Population Genetics, GlaxoSmithKline Inc., Five Moore Drive, P.O. Box 13398, Research Triangle Park NC 27709, USA

[2] Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203, USA

Corresponding author: Dmitri Zaykin

Abstract

Background

This article describes classical and Bayesian interval estimation of genetic susceptibility based on random samples with pre-specified numbers of unrelated cases and controls.

Results

Frequencies of genotypes in cases and controls can be estimated directly from retrospective case-control data. On the other hand, genetic susceptibility defined as the expected proportion of cases among individuals with a particular genotype depends on the population proportion of cases (prevalence). Given this design, prevalence is an external parameter and hence the susceptibility cannot be estimated based on only the observed data. Interval estimation of susceptibility that can incorporate uncertainty in prevalence values is explored from both classical and Bayesian perspective. Similarity between classical and Bayesian interval estimates in terms of frequentist coverage probabilities for this problem allows an appealing interpretation of classical intervals as bounds for genetic susceptibility. In addition, it is observed that both the asymptotic classical and Bayesian interval estimates have comparable average length. These interval estimates serve as a very good approximation to the "exact" (finite sample) Bayesian interval estimates. Extension from genotypic to allelic susceptibility intervals shows dependency on phenotype-induced deviations from Hardy-Weinberg equilibrium.

Conclusions

The suggested classical and Bayesian interval estimates appear to perform reasonably well. Generally, the use of exact Bayesian interval estimation method is recommended for genetic susceptibility, however the asymptotic classical and approximate Bayesian methods are adequate for sample sizes of at least 50 cases and controls.

Background

Association mapping of complex phenotypes in case-control samples involves analysis of tables of genotype/allele counts collected at a large number of genetic loci. Relating single-locus genotype and allele frequencies to the outcome is a basic analysis step even if complex interactions among loci are expected. Indeed, biologically realistic models that involve multiple interacting polymorphisms may induce considerable "marginal effects" associated with individual loci. Often the numbers of cases and controls are fixed in advance by the experimental design, and the multiple markers are typed. Then case/control proportions remain the same for all markers, whereas genotype and allele numbers in cases and controls are subject to the random sampling variation. The reverse is however of greater interest: what is the genetic susceptibility, or the probability that a random individual will have a particular outcome given that a particular genotype or an allele is observed at a locus? Unlike the odds ratio, this parameter is not invariant with respect to the prospective vs. retrospective sampling schemes. Therefore, the point estimate of this probability can only be obtained from genotype counts in cases and controls by assuming a particular value of the population prevalence. Another issue is the degree of uncertainty in the estimate, which is being investigated in this article via interval estimation. Classical, or "frequentist" confidence interval (CI) is a well established framework for interval estimation. Such an interval is a random quantity, and the probability statements are made about proportions of times a random CI covers the fixed population parameter. A more relevant question is often about the vari-

ability, or uncertainty associated with the estimate of the population value. In other words, we would like to be able to make statements about the lower and upper bounds for genetic susceptibility, thus interpreting the interval as fixed, and the susceptibility as random, where randomness is due to the limited amount of data about the parameter. Bayesian (e.g. "credible") intervals provide such interpretation, but sometimes are criticized for subjectivity associated with the choice of prior distributions. It is not generally the case that classical and Bayesian intervals should correspond to each other, however such correspondence is possible for certain combinations of the likelihood and prior distributions. For example, P. Altham [1] showed that when testing for the difference of two binomial proportions, frequentist Fisher's exact test can be viewed as a Bayesian test when assuming binomial likelihoods and Beta priors for the distributions of proportions. Thus, although no priors are explicitly assumed by Fisher's test, one can recover prior distributions that are indirectly implied by the test. Datta and Mukerjee [2] provide an extensive review of such matching probability problems and show that in most parametric cases suitable priors can be constructed that would match the Bayesian credible intervals of a given size to that of a frequentist confidence interval up to second order. As we show here, under a beta-binomial model there is a close connection between the two types of intervals for genetic susceptibility, which allows flexibility in the interpretation.

Results

Susceptibility intervals

Association studies are often "retrospective" in the sense that samples of cases and controls are determined by recruitment ("fixed"), and the genetic variants, e.g. genotypes AA and $\overline{AA}$ (not-AA) at a genetic marker, are random. Estimated genotype frequencies among cases or controls, e.g. $\widehat{\Pr}(AA \mid \mathcal{Y})$ are obtained directly from the following table as $\widehat{\Pr}(AA \mid \mathcal{Y}) = x/n$.

	AA	$\overline{AA}$	
Cases ($\mathcal{Y}$)	$n_{11} = x$	$n_{12} = n - x$	
Controls ($\mathcal{N}$)	$n_{21} = y$	$n_{22} = m - y$	

Assume AA is the risk genotype. If the samples were random with respect to both rows and columns, then the estimated susceptibility (penetrance, or the positive predictive value), $\phi = \Pr(\mathcal{Y} \mid AA)$, would simply be estimated by $x/(x + y)$. Because the rows are fixed, we need the population prevalence, w. Then,

$$\begin{aligned} \phi &= \Pr(\mathcal{Y} \mid AA) = \frac{w \Pr(AA \mid \mathcal{Y})}{w \Pr(AA \mid \mathcal{Y}) + (1-w) \Pr(AA \mid \mathcal{N})} \\ &= \frac{wp}{wp + (1-w)q} \\ &= \left[1 + \frac{1-w}{w} \frac{q}{p}\right]^{-1} \end{aligned} \tag{1}$$

where $w = \Pr(\mathcal{Y})$, $p = \Pr(AA \mid \mathcal{Y})$, and $q = \Pr(AA \mid \mathcal{N})$. If the rows are assumed to arise independently from binomial distributions, the maximum likelihood estimates of the parameters p and q are given by $\hat{p} = x/n$, and $\hat{q} = y/m$, respectively. The population preva-

lence, w, has to be estimated externally, rather than from the data provided by retrospective case-control samples. Let $\hat{w}$ denote such an estimate of w. Hence, a point estimate of ϕ can be obtained as $\hat{\phi} = \left[1 + \frac{1-\hat{w}}{\hat{w}} \ \frac{\hat{q}}{\hat{p}}\right]^{-1} = \left[1 + \frac{1-\hat{w}}{\hat{w}} \ \frac{ny}{mx}\right]^{-1}$.

The expression $\hat{\phi}$ given above involves the ratio of random variables, y/x, and hence obtaining the exact sampling variance is problematic. Approximate methods have been suggested for the ratio of binomial variables [3, 4]. An approximate variance using the logit transformation allows us to obtain a simple expression as well as to separate the terms for p, q and w in the variance expression. To see this, define

$$\begin{aligned} \eta &= \ln \frac{\phi}{1-\phi} \\ &= \ln \frac{w}{1-w} + \ln p - \ln q \end{aligned} \tag{2}$$

Classical interval

The first-order Taylor series approximation gives the variance of $\hat{\eta} = \ln\left[\hat{\phi}/(1-\hat{\phi})\right]$ as

$$V(\hat{\eta}) \approx \frac{V(\hat{p})}{p^2} + \frac{V(\hat{q})}{q^2} + V\left(\ln \frac{\hat{w}}{1-\hat{w}}\right) \tag{3}$$

where the first two terms refer to the variance of the relative risk of AA on the log scale. Pepe [5] presented extensive discussion on estimation of similar quantities in the context of biomedical research. The last term of (3) can be further approximated as

$$V\left(\ln \frac{\hat{w}}{1-\hat{w}}\right) \approx \frac{V(\hat{w})}{(1-w)^2 w^2} \tag{4}$$

and requires the knowledge of $\hat{w}$ which must come from an external source. When the range ($r = w_u - w_l$) of w is known, we may define $r^* = \ln \frac{w_u}{1-w_u} - \ln \frac{w_l}{1-w_l}$, the range on

the logit scale. Then $V(\ln \frac{\hat{w}}{1-\hat{w}})$ can be approximated by $(r^*/6)^2$, since three standard deviations from the mean cover over 99% of the (centered) normal distribution, $f(x, 0, \sigma)$, i.e. $0.99 < 2\int_0^{3\sigma} f(x, 0, \sigma)dx$, where $f(x, 0, \sigma)$ denotes the probability density function of a Normal distribution with mean 0 and standard deviation σ. Henceforth, unless otherwise mentioned we use $\hat{V}\left(\ln \frac{\hat{w}}{1-\hat{w}}\right) = (r^*/6)^2$. Browne [6] discussed this issue and similar approaches of relating the range to the standard deviation.

The CI for susceptibility is obtained by inverting the endpoints of the asymptotic normal interval for η,

$$\hat{\eta} \pm z_{\alpha/2}\sqrt{\hat{V}(\hat{\eta})} \tag{5}$$

e.g. the upper point $u = \hat{\eta} + z_{\alpha/2}\sqrt{\hat{V}(\hat{\eta})}$ is inverted as $\frac{e^u}{1+e^u}$. As we know that $\hat{V}(\hat{p}) = \hat{p}(1-\hat{p})/n$, and $\hat{V}(\hat{q}) = \hat{q}(1-\hat{q})/m$, the estimated variance becomes

$$\hat{V}(\hat{\eta}) \approx \frac{1}{x} - \frac{1}{n} + \frac{1}{y} - \frac{1}{m} + \hat{V}\left(\ln \frac{\hat{w}}{1-\hat{w}}\right) \tag{6}$$

Approximate Bayesian interval

The variance given by (6) is infinite when the observed value of either x or y is zero. Before describing a method of dealing with this, we note that the usual asymptotic variance formula for the log odds ratios, based on the sum of reciprocals of the 2×2 table counts, $\{n_{11}, n_{12}, n_{21}, n_{22}\}$, has a similar deficiency, i.e. it results in an infinite variance when one of the observations is zero. To avoid this problem, it is common to add 1/2 to each cell. Haldane [7], Gart and Zweifel [8] justified this on the basis of minimizing the bias in the

estimation of the logit variance (also see discussion in [9]). The variance becomes

$$\frac{1}{n_{11}+1/2}+\frac{1}{n_{12}+1/2}+\frac{1}{n_{21}+1/2}+\frac{1}{n_{22}+1/2} \tag{7}$$

We propose a similar modification of (6) and justify it by showing that it is an approximate Bayesian variance estimator. In passing, we note that (7) can also be obtained by using the same argument as in the derivation of the approximate posterior variance of $\hat{\eta}$ that follows.

When the sampling distribution of genotypes is binomial, i.e. $x|p \sim \text{Bin}(p, n)$ and $y|q \sim \text{Bin}(q, m)$ and the prior distributions for p and q are independent Beta(γ_1, β_1) and Beta(γ_2, β_2) respectively, the posterior distributions of p and q are independent Beta's with

$$\begin{aligned} p \mid x &\sim \text{Beta}(x+\gamma_1, n-x+\beta_1) \\ q \mid y &\sim \text{Beta}(y+\gamma_2, m-y+\beta_2) \end{aligned} \tag{8}$$

The binomial likelihood arises as a consequence of independence between genotypes in a random sample. Although a Beta prior for allele frequencies is sometimes justified from purely mathematical convenience, it does provide a sufficiently flexible shape. More importantly, certain population-genetic models result in Beta, or more generally Dirichlet distributions for allele frequencies. For example, population frequencies at mutation-drift equilibrium follow this distribution [10]. A reasonable assumption is that a typical value from the susceptibility distribution is around the prevalence value, which implies similarity in the prior distributions of p and q, i.e. $\gamma_1 = \gamma_2, \beta_1 = \beta_2$.

Different parameters would generally assume prior deviation of the typical susceptibility value away from the population prevalence.

Posterior expectations and variances of p and q are respectively given by,

$$\begin{aligned} E(p \mid x) &= \frac{x+\gamma_1}{n+\gamma_1+\beta_2} \\ E(q \mid y) &= \frac{y+\gamma_2}{m+\gamma_2+\beta_2} \\ V(p \mid x) &= \frac{E(p \mid x)\,(1-E(p \mid x))}{n+1+\gamma_1+\beta_1} \\ V(q \mid y) &= \frac{E(q \mid y)\,(1-E(q \mid y))}{m+1+\gamma_2+\beta_2} \end{aligned} \tag{9}$$

Using posterior variances for $\hat{V}(\hat{p})$ and $\hat{V}(\hat{q})$, and substituting posterior expectations for parameter estimates $\hat{p}$ and $\hat{q}$ in (3), we obtain

$$\begin{aligned} \hat{V}(\hat{\eta} \mid \gamma_1, \gamma_2, \beta_1, \beta_2) &\approx \frac{n+\gamma_1+\beta_1}{n+\gamma_1+\beta_1+1}\left(\frac{1}{x+\gamma_1}-\frac{1}{n+\gamma_1+\beta_1}\right) \\ &+ \frac{m+\gamma_2+\beta_2}{m+\gamma_2+\beta_2+1}\left(\frac{1}{y+\gamma_2}-\frac{1}{m+\gamma_2+\beta_2}\right) \\ &+ \hat{V}\left(\ln\frac{\hat{w}}{1-\hat{w}}\right) \end{aligned} \tag{10}$$

Again, the approximation (4) is appropriate when $\hat{V}(\hat{w})$ is available.

When $\gamma_1, \gamma_2 \to 0$ and $\beta_1, \beta_2 \to 0$ (Haldane's prior, see [11]) we are essentially back to (6),

$$\begin{aligned} \hat{V}(\hat{\eta} \mid 0,0,0,0) &\approx \frac{n}{n+1}\left(\frac{1}{x}-\frac{1}{n}\right)+\frac{m}{m+1}\left(\frac{1}{y}-\frac{1}{m}\right) \\ &+ \hat{V}\left(\ln\frac{\hat{w}}{1-\hat{w}}\right) \end{aligned} \tag{11}$$

because $n/(n+1)$ and $m/(m+1)$ approach one as m and n increase. Therefore (6) is also justified from the Bayesian point of view.

When $\gamma_1 = \gamma_2 = \beta_1 = \beta_2 = 1/2$ (Jeffreys' prior, [11]),

$$\hat{V}(\hat{\eta} \mid \tfrac{1}{2},\tfrac{1}{2},\tfrac{1}{2},\tfrac{1}{2}) \approx \frac{n+1}{n+2}\left(\frac{1}{x+\frac{1}{2}}-\frac{1}{n+1}\right)+\frac{m+1}{m+2}\left(\frac{1}{y+\frac{1}{2}}-\frac{1}{m+1}\right)$$

$$+ \quad \hat{V}\left(\ln\frac{\hat{w}}{1-\hat{w}}\right) \tag{12}$$

Welch and Peers [12] established that in general Jeffreys' priors provide mathematical correspondence between frequentist and Bayesian intervals. In particular, these authors proved that under some regularity conditions (which are trivially satisfied for this problem) Jeffreys' prior is the unique prior for a parameter for which a Bayesian credible interval and a frequentist confidence interval match up to the second order.

The asymptotic normal credible interval based on $\tilde{\eta} = \ln\left[\tilde{\phi}/(1-\tilde{\phi})\right]$ can be obtained as:

$$\tilde{\eta} \pm z_{\alpha/2}\sqrt{\hat{V}(\tilde{\eta})} \tag{13}$$

where $\tilde{\phi} = \frac{\hat{w}E(p|x)}{\hat{w}E(p|x)+(1-\hat{w})E(q|y)}$, and the endpoints, l and u, are inverted as $\frac{e^u}{1+e^u}$ and $\frac{e^l}{1+e^l}$, respectively to produce an approximate $(1-\alpha)\%$ interval estimate of ϕ.

Exact Bayesian interval

It is also possible to obtain an "exact" Bayesian credible interval using Monte Carlo samples generated from the posterior distribution of ϕ. The sample is obtained by repeatedly drawing p and q from their posterior distributions given by (8) and calculating ϕ via (1) for each realization of p, q. This generates an empirical posterior distribution for ϕ and a $(1-\alpha)\%$ interval is given by $(\alpha/2)$, $(1-\alpha/2)$ quantiles of this distribution. Values w can be drawn uniformly from the range of its possible values. A bell-shaped Beta distribution can also be assumed with parameters based on the reported range, $w_l - w_u$, as described in Methods, section 3. An algorithm is as follows:

1. Generate $p^{(i)} \sim \text{Beta}(x+\gamma_1, n-x+\beta_1)$, $i = 1, .., B$.

2. Generate $q^{(i)} \sim \text{Beta}(y+\gamma_2, m-y+\beta_2)$, $i = 1, .., B$.

3. Generate $w^{(i)} \sim \text{Uniform}(w_l, w_u)$, $i = 1, .., B$.

4. Compute $\phi^{(i)} = \frac{w^{(i)}p^{(i)}}{w^{(i)}p^{(i)}+(1-w^{(i)})q^{(i)}}$, $i = 1, ..., B$.

Then the $\alpha/2$ and $1-\alpha/2$ quantiles of the generated empirical posterior distribution of ϕ provide the $(1-\alpha)\%$ credible interval for the susceptibility. This algorithm is easily implemented in R, a language and environment for statistical computing and graphics available as Free Software at `http://www.r-project.org/` . For example, setting $\gamma_1 = \gamma_2 = \beta_1 = \beta_2 = 0.5$, the entire code using 100,000 samples, 95% credible interval, and w distributed uniformly between 0.04 and 0.06 is

```
p <- rbeta(100000, x+0.5, n-x+0.5)
q <- rbeta(100000, y+0.5, m-y+0.5)
w <- runif(100000, 0.04, 0.06)
phi <- w*p / (w*p + (1-w)*q)
phiL <- quantile(phi, prob=0.025)
phiU <- quantile(phi, prob=0.975)
cat(c("Interval estimate", phiL, phiU), fill=T)
```

Common beta priors on p and q ($\gamma_1 = \gamma_2, \beta_1 = \beta_2$) may be interpreted as vague with respect to the prior distribution for ϕ. On the logit scale, this distribution is bell-shaped symmetric, and centered around $\ln[w/(1-w)]$. Its variance decreases as the common (γ, β) parameters for p and q increase. For large values of γ, β this variance can be obtained by the Taylor series approximation, $2\beta/[\gamma(\gamma+\beta+1)]$. A mixture of beta distributions can take an arbitrary shape with the great advantage that the resulting distribution is bounded between

the lower and upper values – which is the case for p, q, and ϕ parameters. In Methods (section 1) we outline the sampling scheme for the mixture.

Interval estimation of allele or haplotype susceptibilities

Methods described so far could be applied without modifications to the estimation of allele or haplotype susceptibilities when the probability of obtaining a sample of alleles follows the multinomial likelihood. However, such usage would require the assumption of independence between alleles. This can be justified under the assumptions of Hardy-Weinberg equilibrium (HWE) in the population, as well as the multiplicative effects of haplotype susceptibilities [13]. When these assumptions are reasonable, the intervals are obtained in the same way as described above using counts of alleles/haplotypes instead of genotypes. Otherwise the variance of allele frequencies, $\hat{p}_A = x/n$, $\hat{q}_A = y/m$ is no longer binomial (x, y are now the counts of alleles A in cases and controls respectively, and n, m are the total numbers of alleles). It is instead given by

$$\begin{aligned} V(\hat{p_A}) &= 1/n\,[p_A(1-p_A) + D_p] \\ V(\hat{q_A}) &= 1/m\,[q_A(1-q_A) + D_q] \end{aligned} \tag{14}$$

where D_p, D_q are the Hardy-Weinberg disequilibrium (HWD) coefficients [14], defined as the deviation of the observed frequencies of AA from the expected under random union of alleles. For example, $D_p = \Pr(AA \mid \mathcal{Y}) - p_A^2$, and is estimated by using the observed frequencies, $\hat{D}_p = \hat{\Pr}(AA \mid \mathcal{Y}) - \hat{p}_A^2$. These variances are most easily derived by re-coding the genotypes

$(AA, A\overline{A}, \overline{AA})$ as $G = (-1, 0, 1)$, which keeps track of the number of alleles A minus one in a corresponding genotype, and taking expectations in $V(G) = E(G^2) - (E(G))^2$ (see Appendix 1 in [15], or two alternative derivations in [14]).

These coefficients (quantifying HWD among cases or controls) may be non-zero even if the population is in HWE [13]. Then the variance (6) becomes

$$\hat{V}_D(\hat{\eta}) \approx \frac{1}{x} - \frac{1}{n} + \frac{1}{y} - \frac{1}{m} + \hat{V}\left(\ln \frac{\hat{w}}{1-\hat{w}}\right) + \hat{D}_p \frac{n}{x^2} + \hat{D}_q \frac{m}{y^2} \tag{15}$$

with two new additional terms to account for HWD. Consider the effect of $\hat{D}_p$. If $\hat{D}_p$ is negative, the resulting variance becomes smaller. If it is positive, the maximum possible value of $\hat{D}_p$ [14] is

$$\begin{aligned} \max(\hat{D}_p) &= \hat{p_A}(1 - \hat{p_A}) \\ &= \frac{x}{n}\left(1 - \frac{x}{n}\right) \end{aligned} \tag{16}$$

The maximum values for the terms with HWD in (15) are

$$\begin{aligned} \hat{D}_p \frac{n}{x^2} &= \frac{1}{x} - \frac{1}{n} \\ \hat{D}_q \frac{m}{y^2} &= \frac{1}{y} - \frac{1}{m} \end{aligned} \tag{17}$$

Therefore, when $\hat{D}_p$ or $\hat{D}_q$ reach their maximum possible values, the inflation of the corresponding part of the variance is twice the value assuming the equilibrium. To set up an exact Bayesian interval for allelic/haplotypic susceptibility we may start with distributions for genotypes/diplotypes instead of that for alleles/haplotypes. Random samples of genotypes for cases, $\{n_{1j}\}$, and for controls, $\{n_{2j}\}$, follow the multinomial distribution. A

conjugate prior distribution is Dirichlet(γ_{ij}), so the posterior genotype frequencies P_{ij} are sampled from Dirichlet($\gamma_{ij} + n_{ij}$) distribution. Posterior allele/haplotype frequencies are given as $p_i = P_{ii} + \sum_{i \neq j} P_{ij}/2$ and the method proceeds as above. Jeffrey's prior is obtained when all γ_{ij} are 1/2.

Coverage probabilities

The intervals using (6, 10) and the exact Bayesian interval have good frequentist coverage probabilities. To illustrate this, we simulated population values of p and q as $p \sim \text{Beta}(1, 1)$ and $q \sim \text{Beta}(1, 1)$ for each of 10,000 simulation runs. We used $\gamma = \beta = 1/2$ for the interval calculations and assumed that $w = 0.5$ or $w = 0.04$ is known and fixed. To compare intervals based on (6) and (12) we only considered samples with $x, y > 0$, obtained as $x \sim \text{Bin}(p, n)$, $y \sim \text{Bin}(q, m)$. In addition, we considered binomial samples with at least one of x, y equal to zero to check coverage properties of (12). We used $m = n = 10, 50, 100, 500$ and 90% coverage intervals. Frequentist intervals cannot be used with zero counts of n_{11} or n_{21}, which may be problematic only for sample sizes of 10 (17% of the samples). In this small percentage of cases, we resort to adding 1/2 to the counts as in (10). Agresti [9] considered a similar approach for calculation of CIs for odds ratios, that is to use Gart's formula (7) only when zero counts are encountered. Results for coverage probabilities are shown in Table 1. Note that the coverage of all three methods is around the nominal 90% for all sample sizes and both values of population prevalence. Results in Table 1 are averages across the distribution of p and q, however similar coverage results are obtained when the population values of p, q

are set to specific fixed values. We considered the following fixed sets of parameters:

1. $p = 0.96, q = 0.06, w = 0.04$. This set results (equation 1) in the population susceptibility value of $\phi = 0.4$ (ten-fold increase from the population prevalence value).

2. $p = 0.9, q = 0.15, w = 0.04$. This set results in the population susceptibility value of $\phi = 0.2$.

3. $p = 0.8, q = 0.2, w = 0.5$. This set results in the population susceptibility value of $\phi = 0.8$.

The coverage probabilities for all three methods were found to be sufficiently close to the nominal values of 90% and 95% for sample sizes of 50 or larger (Table 2).

Table 3 shows the mean length and the standard deviation of the interval length across 10,000 simulations used to produce Table 1. This standard deviation reflects estimated variability of the interval length and decreases with sample size, but not with the number of simulations. The corresponding standard error could be obtained by dividing the standard deviation by the square root of the number of simulations. Table 4 shows the same results among the intervals that include the population value of w. Numbers for each parameter combination that contrast three intervals in a given table are obtained using the same data sets. For example, once a sample of size n is obtained for a specific value of w, all three intervals are calculated using this sample. Therefore, comparisons between methods are quite precise given the number of simulations. Again, the mean and standard deviation are similar among three methods. Nevertheless, the classical CI is never the shortest of the three

intervals for the parameter values that we have considered and typically it has the highest length variability. Using data summarized in Table 3, the exact interval was found to have the smallest length 5 times out of 8 considered parameter combinations and has the lowest variability 5 times out of 8. These numbers are respectively 5 out of 8 and 4 out of 8 for data used to construct Table 4. There is some dependency between results from Tables 3 and 4 due to the same combination of parameters.

We do not present numerical results based on mis-specification of w or the uncertainty in it, because the behavior of intervals is clear. When w is mis-specified, the intervals actively worsen as the sample size increases (in the sense of the probability of including the population parameter), since they shrink around the wrong value. On the other hand, the uncertainty in w widens the interval length. As the sample size becomes large, the minimum length simply reflects the variability in w. We conducted simulation experiments to illustrate this point, using the population value of $w = 0.045$ assuming the range 0.03 to 0.06 is reported. For the exact Bayesian interval, the distribution of w was modelled as Beta(ψ_1=77.31, ψ_2=1640.69) with parameters estimated by (29). These values are taken to reflect the reported prevalence of the hypersensitivity reaction to an antiretroviral drug abacavir, for which genetic predisposition has been described [16]. Table 5 reports results for the coverage probabilities and the interval length of 10% intervals, assuming $p = 0.8$ and $q = 0.2$, which corresponds to the population susceptibility value of 0.159. The results show that the average length is somewhat increased when the calculation are based on the range of w, and the interval coverage is above the nominal value of 90%, because of the increased

interval width.

Pharmacogenetic application

Hypersensitivity reaction to the antiretroviral drug abacavir affects approximately 4.5% of patients. The hypersensitivity symptoms are varied with fever and rash among the most common. A small number of fatalities have also been reported. The exact mechanism of hypersensitivity is not established, but the pattern of symptoms suggests that it is an immunological reaction, triggered by specific genetic polymorphisms. Hetherington et al. [16] studied the association of HLA-B57 haplotype with hypersensitivity. One or two copies of HLA-B57 haplotype were present in $x = 39$ of $n = 84$ individuals with hypersensitivity (cases) and in $y = 4$ out of $m = 113$ controls. The prevalence of the hypersensitivity reaction is estimated to be between 0.03 and 0.06 [16]. The estimate of ϕ using (1) and the midrange w (that is equal to 0.045) is 0.382. Assuming $w = 0.045$ is fixed (i.e. $V(w) = 0$), the intervals are as follows. The frequentist 95% interval is 0.187 – 0.624, the asymptotic Bayesian interval is 0.180 – 0.584, and the exact Bayesian interval is 0.203 – 0.648. Taking the range of w into account, the frequentist and Bayesian asymptotic intervals become wider, 0.183 – 0.631 and 0.175 – 0.591, correspondingly, and the exact Bayesian interval assuming the uniformity of w on the range (0.03 – 0.06) is 0.186 – 0.658. This interval is estimated from repeatedly $(i = 1, ..., 100000)$ generating $p^{(i)} \sim \text{Beta}(39+1/2, 84-39+1/2)$, $q^{(i)} \sim \text{Beta}(4+1/2, 113-4+1/2)$, $w^{(i)} \sim \text{Uniform}(0.03, 0.06)$, and calculating $\phi^{(i)} = \frac{w^{(i)}p^{(i)}}{w^{(i)}p^{(i)}+(1-w^{(i)})q^{(i)}}$. Then the credible interval was obtained from this (empirical) distribution of ϕ by determining its

2.5% and 97.5% quantiles. It is not necessary to assume the uniformity of w. Indeed, it may be reasonable to assume that the midrange value is more plausible than the endpoints and a realistic distribution can be described as bell-shaped. Taking $w \sim$ Beta(57.8572,1227.86) generates a bell-shaped distribution with the mean at 0.045 and 99% of this distribution is contained between 0.03 and 0.06 (Methods, section 3). The resulting interval is somewhat smaller, 0.195 – 0.653. Results for this example are summarized in Table 6.

Discussion

Asymptotic frequentist and Bayesian intervals for genetic susceptibility as well as the exact Bayesian interval are quite similar in properties. Given that the sampling from the posterior distribution can be done directly, there is no particular reason to resort to approximations and the exact Bayesian interval can be recommended. Nevertheless, approximations are useful in that they reveal connections between the confidence and credible intervals as well as allow for sample size and power calculations as we outline in Methods, section 2. The mean length and the interval variability appear somewhat smaller for the Bayesian intervals. The frequentist coverage of Bayesian intervals is satisfactory. One may even argue that the operational usage of CIs is more often Bayesian than it is not, because the practical issue is ultimately about the confidence on the plausible range of parameter values. Algebraic relations between classical and Bayesian intervals as well as similarities in numerical properties allow Bayesian interpretation for the CIs. On the other hand, Bayesian intervals for this problem can be described in the classical sense as random, covering the fixed population parameter $\alpha\%$ of the time in repeated samples.

Caution should be taken when the inference is about allele or haplotype susceptibilities. When assumption of population HWE does not hold, or when there are substantial deviations from multiplicativity, the variance of susceptibility contains additional terms that account for population or model-induced deviations from Hardy-Weinberg equilibrium.

Conclusions

We found that the classical interval for genetic susceptibility can be considered as an approximate Bayesian interval under the beta-binomial model. This algebraic similarity between the proposed classical and approximate Bayesian intervals allows an appealing Bayesian interpretation of the usual confidence limits. Simulation studies also confirm similarities in coverage probabilities and interval lengths among asymptotic classical and approximate and exact Bayesian intervals.

Methods

1. Posterior sampling using the mixture of beta distributions

We derive explicitly the posterior distribution of θ under a Binomial sampling using a mixture of Beta priors. As Beta is a conjugate prior for the binomial likelihood, we show that mixture of Beta's is also conjugate. To see this, consider the problem

$$X \mid \theta \sim \text{Bin}(\theta, n) \tag{18}$$

$$\Rightarrow p(x \mid \theta) \propto \theta^x (1-\theta)^{n-x} \tag{19}$$

where $p(x \mid \theta)$ denotes the sampling density of $X = x$ given θ. Now consider a prior for θ given by a mixture of k beta distributions $\text{Beta}(\gamma_j, \beta_j)$ with weights v_j, such that $\sum_{j=1}^{k} v_j = 1$ and $v_j \geq 0$ for all j. More explicitly, the prior density $\pi(\theta)$, of θ is given by,

$$\pi(\theta) = \sum_{j=1}^{k} v_j b(\theta \mid \gamma_j, \beta_j) \tag{20}$$

$$\text{i.e., } \theta \sim \sum_{j=1}^{k} v_j \text{Beta}(\gamma_j, \beta_j) \tag{21}$$

where $b(\theta \mid \gamma_j, \beta_j) = \theta^{\gamma_j - 1}(1-\theta)^{\beta_j - 1}/B(\gamma_j, \beta_j)$ denotes the density of a $\text{Beta}(\gamma_j, \beta_j)$ distribution and $B(\gamma_j, \beta_j) = \int_0^1 \theta^{\gamma_j - 1}(1-\theta)^{\beta_j - 1} d\theta$ denotes the Beta function. Using Bayes' rule it follows that the posterior distribution is given by,

$$\pi(\theta \mid x) = \frac{\sum_{j=1}^{k} v_j \theta^{\gamma_j + x - 1}(1-\theta)^{\beta_j + n - x - 1}}{\sum_{j=1}^{k} v_j \text{B}(\gamma_j + x, \beta_j + n - x)} \tag{22}$$

$$= \sum_{j=1}^{k} v_j^* \; b(\theta \mid \gamma_j^*, \beta_j^*) \tag{23}$$

where $\gamma_j^* = \gamma_j + x,\ \beta_j^* = \beta_j + n - x$, and

$$v_j^* = \frac{v_j\ \mathrm{B}(\gamma_j - x, \beta_j + n - x)}{\sum_{j=1}^k v_j\ \mathrm{B}(\gamma_j - x, \beta_j + n - x)}$$

Thus it follows that if $X|\theta \sim \mathrm{Bin}(\theta, n)$ and $\theta \sim \sum_j v_j \mathrm{Beta}(\gamma_j, \beta_j)$ then $\theta|X = x \sim \sum_j v_j^* \mathrm{Beta}(\gamma_j^*, \beta_j^*)$. It may be noted that any continuous density on $[0, 1]$ can be approximated by an appropriate mixture of Beta densities (see [17] for more general results on approximating any prior using natural conjugate priors). Besides such theoretical properties, sampling from the mixture of Beta's is also straightforward. For instance, in order to generate $\theta \sim \sum_j v_j^* \mathrm{Beta}(\gamma_j^*, \beta_j^*)$, we can apply the following two-stage sampling:

1. Sample $J = j \in \{1, ..., k\}$ with probability v_j^*, i.e., $\Pr[J = j] = v_j^*$.

2. Conditional on the sampled value of $J = j$, sample $\theta \sim \mathrm{Beta}(\gamma_j^*, \beta_j^*)$.

This generates a sample from the mixture $\sum_j v_j^* \mathrm{Beta}(\gamma_j^*, \beta_j^*)$. Thus, our sampling scheme described in "Exact Bayesian Interval" section can be easily modified to sampling from mixture of Beta's instead.

2. Power and sample size

Under the hypothesis of equality of genotype frequencies in cases and controls, $H_0 : \Pr(AA \mid \mathcal{Y}) = \Pr(AA \mid \mathcal{N})$, the expected susceptibility in (1) is equal to the population prevalence, w. Correspondingly, the logit transformation η has the mean equal to $\ln \frac{w}{1-w}$, and

$$\frac{\hat{\eta} - \ln \frac{w}{1-w}}{\sqrt{\frac{1}{x} - \frac{1}{n} + \frac{1}{y} - \frac{1}{m} + (r^*/6)^2}} \sim N(0,1) \tag{24}$$

The two-sided alternative hypothesis is $H_A : \Pr(AA \mid \mathcal{Y}) \neq \Pr(AA \mid \mathcal{N})$. The values of $\Pr(AA \mid \mathcal{Y})$, $\Pr(AA \mid \mathcal{N})$ and the ratio $\frac{\Pr(AA|\mathcal{Y})}{\Pr(AA|\mathcal{N})}$ under the alternative are denoted as p^A, q^A, and Δ^A, respectively. Once these values are specified, we can calculate the power of detecting the difference of the susceptibility from the population prevalence as well as the sample size required to achieve the required power under an acceptable type I error rate. We will illustrate the sample size calculation assuming the equal number of cases and controls ($m = n$). Power calculation given the fixed sample size is obtained similarly. We have

$$\frac{\hat{\eta} - \ln \frac{w}{1-w}}{\sqrt{\hat{V}(\hat{\eta})}} \sim N(\delta, 1) \tag{25}$$

where

$$\begin{aligned} \delta &= \frac{\ln(\Delta^A)}{\sqrt{\frac{1}{n}(\frac{1}{\Delta^A q^A} + \frac{1}{q^A} - 2) + \frac{\hat{V}(\hat{w})}{(1-\hat{w})^2 \hat{w}^2}}} \\ &= Z_{\alpha/2} + Z_\beta \end{aligned} \tag{26}$$

is the non-centrality parameter corresponding to $1-\beta$ power under a two-sided level-α test. Therefore, the necessary sample size is

$$n = \frac{\frac{1}{\Delta^A q^A} + \frac{1}{q^A} - 2}{\left(\frac{\ln(\Delta^A)}{Z_{\alpha/2}+Z_\beta}\right)^2 - \frac{\hat{V}(\hat{w})}{(1-\hat{w})^2\hat{w}^2}} \tag{27}$$

3. Estimating parameters of the Beta distribution for the prevalence from the range of reported values

Suppose the range of the prevalence values, $w_l - w_u$, is reported. We would like to model the uncertainty with a bell-shaped Beta($\psi_1 > 1, \psi_2 > 1$) distribution, such that a certain percentage of the distribution (e.g. 99%) is between w_l and w_u with the expected value given by $w_l + (w_u - w_l)/2$. The parameters of this distribution can be found as the solution for ψ_1 of

$$0.99 = \int_{w_l}^{w_u} b\left(\theta \mid \psi_1, \psi_1(1-\hat{w})/\hat{w}\right) d\theta \tag{28}$$

where $b(\theta \mid \psi_1, \psi_2) = \theta^{\psi_1 - 1}(1-\theta)^{\psi_2 - 1}/B(\psi_1, \psi_2)$ denotes the density of a Beta(ψ_1, ψ_2) distribution and $B(\psi_1, \psi_2) = \int_0^1 \theta^{\psi_1 - 1}(1-\theta)^{\psi_2 - 1} d\theta$ is the Beta function. The second parameter for $b(\cdot)$ in (28) is given by the constraint that the expectation for w is $\psi_1/(\psi_1 + \psi_2)$.

As a sample calculation, consider an antiretroviral drug abacavir for which the hypersensitivity reaction range is reported as $w_l = 0.03$ to $w_u = 0.06$ [16], so that we can take $\hat{w} = w_l + (w_u - w_l)/2 = 0.045$. If we assume that these bounds cover 99% of the prevalence distribution that results from sampling errors in estimation, possibly confounded with other factors such as genuine population heterogeneity of samples, then the resulting distribution reflecting the uncertainty in w is Beta(57.8572, 1227.86), using (28).

Much more simply calculated parameter estimates can be obtained by the normal approximation to the beta distribution. Let $\hat{w} = w_l + (w_u - w_l)/2$ and approximate the variance by $\hat{\sigma}^2 = [(w_u - w_l)/6]^2$. By the method of moments, estimates of the parameters (ψ_1, ψ_2)

are

$$\begin{aligned} \hat{\psi}_1 &= \hat{w}\left[(1-\hat{w})/\hat{\sigma}^2 - 1\right] \\ \hat{\psi}_2 &= (1-\hat{w})\left[(1-\hat{w})/\hat{\sigma}^2 - 1\right] \end{aligned} \tag{29}$$

For the abacavir example, these estimates are found as $\hat{\psi}_1 = 77.31$ and $\hat{\psi}_2 = 1640.69$. Using these estimates, 99.7% of the resulting Beta(77.31, 1640.69) distribution is contained within the range 0.03 – 0.06. Thus, the normal approximation yields a somewhat more condensed distribution.

Authors' contributions

All authors made substantial contributions to this paper, including conceiving of the ideas, discussion and writing. All authors read and approved the final manuscript.

Acknowledgements

Clive Bowman and three anonymous reviewers provided valuable comments that improved the manuscript.

References

1. Altham P: **Exact Bayesian analysis of a 2×2 contingency table, and Fisher's "exact" significance test**. J. of the Royal Statistical Society, Series B 1969, **31**: 261–269

2. Datta, GS Mukerjee R: *Probability matching priors: higher order asymptotics*. Lecture Notes in Statistics. Springer Verlag; 2004

3. Katz D, Baptista J,Azen SP, Pike MC: **Obtaining confidence intervals for the risk ratio in cohort studies**. Biometrics 1978, **34**: 469–474

4. Koopman PAR: **Confidence intervals for the ratio of two binomial proportions**. Biometrics 1984, **40**: 513–517

5. Pepe MS: *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press; 2003

6. Browne RH: **Using the sample range as a basis for calculating sample size in power calculations**. Am Statist 2001, **55**: 293–298

7. Haldane JBS: **The estimation and significance of the logarithm of a ratio of**

frequencies. Ann Human Genet 1955, **20**: 309–311

8. Gart JJ, Zweiful JR: **On the bias of various estimators of the logit and its variance with applications to quantal bioassay**. Biometrika 1967, **54**: 181–187

9. Agresti A: **On logit confidence intervals for the odds ratio with small samples**. Biometrics 1999, **55**: 597–602

10. Wright S: **The genetical structure of populations**. Ann Eugen 1951, **15**: 32–354

11. Geisser S: **On prior distributions for binary trials**. The American Statistician 1984, **38**: 244–247

12. Welch BL, Peers, HW: **On formulae for confidence points based on integrals of weighted likelihoods**. J of the Royal Stat Soc (B) 1963, **25**: 318–329

13. Nielsen DM, Ehm MG, Weir BS: **Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus**. Am J Human Genet 1999, **63**: 1531–1540

14. Weir BS: *Genetic data analysis II*. Sinauer, Sunderland MA; 1996

15. Meng Z, Zaykin DV, Xu C-F, Wagner M, Ehm MG: **Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes**. Am J Human Genet 2003, **73**: 115–130.

16. Hetherington S, Hughes AR, Mosteller M, Shortino D, Baker KL, Spreen W, Lai E, Davies K, Handley A, Dow DJ, Fling ME, Stocum M, Bowman C, Thurmond LM, Roses AD: **Genetic variations in HLA-B region and hypersensitivity reactions to abacavir**. Lancet 2002, **359**: 1121–1122

17. Dalal, SR, and Hall, WJ: **Approximating priors by mixtures of natural conjugate priors**. J of the Royal Stat Soc (B) 1983, **45**: 278–286

Table 1. Coverage probabilities for nominal 90% intervals based on 10,000 simulations.

Sample size	10	50	100	500
Prevalence				
	Asymptotic Frequentist			
0.04	0.921	0.910	0.908	0.898
0.50	0.918	0.906	0.906	0.898
	Asymptotic Bayesian			
0.04	0.902	0.903	0.904	0.897
0.50	0.896	0.898	0.901	0.898
	Exact Bayesian			
0.04	0.909	0.902	0.902	0.896
0.50	0.907	0.895	0.901	0.895

Table 2. Coverage probabilities for nominal (90%/95%) intervals based on 10,000 simulations, based on three population settings of (p,q,w). Setting 1: $p = 0.96, q = 0.06, w = 0.04$. Setting 2: $p = 0.9, q = 0.15, w = 0.04$. Setting 3: $p = 0.8, q = 0.2, w = 0.5$.

Sample size	10	50	100	500
(p,q,w) settings				
	Asymptotic Frequentist			
1	0.888/0.951	0.919/0.962	0.916/0.960	0.896/0.948
2	0.942/0.883	0.906/0.962	0.909/0.954	0.896/0.953
3	0.909/0.954	0.908/0.954	0.900/0.947	0.899/0.951
	Asymptotic Bayesian			
1	0.888/0.883	0.918/0.920	0.907/0.951	0.901/0.948
2	0.866/0.920	0.900/0.950	0.904/0.946	0.894/0.951
3	0.879/0.907	0.901/0.950	0.895/0.944	0.899/0.948
	Exact Bayesian			
1	0.911/0.979	0.890/0.922	0.894/0.954	0.897/0.946
2	0.945/0.967	0.900/0.950	0.908/0.950	0.896/0.951
3	0.865/0.954	0.900/0.948	0.896/0.946	0.895/0.950

Table 3. Average length (standard deviation) of the intervals based on 10,000 simulations.

Sample size	10	50	100	500
Prevalence				
		Asymptotic Frequentist		
0.04	0.161(0.220)	0.081(0.151)	0.060(0.121)	0.026(0.057)
0.50	0.361(0.156)	0.168(0.090)	0.117(0.062)	0.052(0.030)
		Asymptotic Bayesian		
0.04	0.143(0.207)	0.075(0.139)	0.057(0.114)	0.025(0.055)
0.50	0.341(0.146)	0.166(0.086)	0.116(0.060)	0.052(0.030)
		Exact Bayesian		
0.04	0.180(0.264)	0.079(0.146)	0.058(0.113)	0.025(0.054)
0.50	0.329(0.152)	0.161(0.085)	0.114(0.059)	0.052(0.029)

Table 4. Average length (standard deviation) of those intervals that contained the true parameter value.

Sample size	10	50	100	500
Prevalence				
		Asymptotic Frequentist		
0.04	0.157(0.215)	0.078(0.147)	0.059(0.117)	0.026(0.056)
0.50	0.364(0.155)	0.168(0.088)	0.119(0.067)	0.052(0.030)
		Asymptotic Bayesian		
0.04	0.139(0.201)	0.072(0.134)	0.056(0.110)	0.026(0.055)
0.50	0.342(0.144)	0.165(0.084)	0.118(0.065)	0.052(0.030)
		Exact Bayesian		
0.04	0.176(0.259)	0.077(0.142)	0.057(0.111)	0.025(0.053)
0.50	0.333(0.150)	0.161(0.084)	0.116(0.065)	0.052(0.030)

Table 5. Effect of uncertainty in the prevalence, w: the population value is $w = 0.045$; the assumed range is either 0.03–0.06 ($\hat{V}(\hat{w}) > 0$, first three columns), or zero ($\hat{V}(\hat{w}) = 0$, last three columns). AF, AB, EB refer to the asymptotic frequentist, approximate Bayesian, and exact Bayesian 10% intervals.

	$\hat{V}(\hat{w}) > 0$			$\hat{V}(\hat{w}) = 0$		
Interval	AF	AB	EB	AF	AB	EB
$n = m = 10$						
Av. Length	0.375	0.308	0.394	0.371	0.303	0.391
Coverage	0.941	0.889	0.868	0.913	0.889	0.868
$n = m = 50$						
Av. Length	0.154	0.144	0.155	0.143	0.134	0.145
Coverage	0.929	0.924	0.920	0.902	0.898	0.895
$n = m = 100$						
Av. Length	0.110	0.107	0.110	0.096	0.093	0.097
Coverage	0.945	0.946	0.939	0.900	0.896	0.900
$n = m = 500$						
Av. Length	0.067	0.067	0.066	0.041	0.041	0.041
Coverage	0.993	0.993	0.992	0.897	0.899	0.900

Table 6. Summary of susceptibility intervals for the pharmacogenetic example. AF, AB, EB refer to the 95% asymptotic frequentist, approximate Bayesian, and exact Bayesian intervals.

	AF	AB	EB
$\hat{V}(\hat{w}) = 0$	0.187-0.624	0.180-0.584	0.203-0.648
$\hat{V}(\hat{w}) > 0$	0.183-0.631	0.175-0.591	0.195-0.653