Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Bowen Shi Date

Exploring Bayesian credible intervals for common epidemiologic effect measures based on cross-sectional data

By

Bowen Shi
Master of Science in Public Health

Biostatistics

_____

Dr. Bob Lyles
Thesis Advisor

_____

Dr. Haber, Michael J
Reader

Exploring Bayesian credible intervals for common epidemiologic effect measures based on cross-sectional data

By

Bowen Shi

B.S.
China Pharmaceutical University
2018

Advisor: Bob Lyles, Ph.D.

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics
2020

Abstract

Exploring Bayesian credible intervals for common epidemiologic effect measures based on cross-sectional data

By Bowen Shi

As the limitations of standard Wald-type methods to estimate the confidence intervals for the difference of proportions, relative risk, and odds ratio in 2 × 2 contingency tables are well recognized, we investigated the frequentist performance of alternative Bayesian credible intervals. We used simulation studies to compare the coverage rates and widths of these competing sorts of confidence intervals. As a new proposal, we also put forth an adjusted credible interval which used two different Dirichlet priors to get the lower and upper limits of the confidence intervals. In small sample settings, this method appears to greatly reduce average interval width compared to Wald-type approaches, while maintaining far better coverage rates compared to more standard credible intervals based on a single Dirichlet prior.

Exploring Bayesian credible intervals for common epidemiologic effect measures based on cross-sectional data

By

Bowen Shi

B.S.
China Pharmaceutical University
2018

Advisor: Bob Lyles, Ph.D.

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics
2020

Table of Contents

1. Introduction

The cross-sectional study is a commonly used method to investigate associations between risk factors and outcomes of interest [1]. The most simple case of a cross-sectional study is to study a binary outcome and a binary risk factor. Consider two independent samples, with X_i, a binomial bin(n_i , p_i) variate. There are three common measures of association that could be estimated in this kind of cross-sectional study based on the resulting two-by-two contingency table. They are odds ratio OR = $[p_1 /(1 – p_1)]/[p_2 /(1 – p_2)]$, the risk difference of proportions RD = $(p_1 – p_2)$ and the relative risk RR = p_1 /p_2. Confidence intervals for these parameters are often of interest to evaluate the association.

The confidence intervals which result from inverting large sample-based Wald tests are the most commonly used. They are often called ‘Wald intervals’ [2], and are known to perform poorly in certain categorical data scenarios when sample size is small [3,4]. Some literature has also proven that their performance could be poor even when the sample size is large, particularly in the case of extreme proportions [5]. To make up for the deficiencies of Wald intervals, several adjustments were proposed in previous literature. Agresti and others suggested adding two ‘success’ and two ‘failures’ before the calculation [6]. Gart suggested

adding 0.5 to each cell of the contingency table when calculating the variance of the estimator [7]. These methods were shown to offer improved performance compared to standard Wald intervals.

Besides adding pseudo observations to the sample, another good direction to perform the adjustment corresponding to shrinkage of point estimates is the Bayesian approach. The main idea of the Bayesian approach is to find a suitable prior distribution for parameters of interest and then to make inferences based on the corresponding posterior distribution. Carlin and Louis considered independent uniform priors for p_1 and p_2 [8]. Their final intervals are similar to Agresti's [6], and can be considered as an approach which adds three 'success' and three 'failures'. In another study [9], beta priors, logit-normal priors, and related correlated priors were simulated and evaluated. Agresti suggested that it is better to use diffuse priors if you want to use a Bayesian estimator and are concerned with frequentist performance.

According to Agresti's results [9] and Brown's suggestion [10], Jeffrey's priors are good options for binomial parameters in a two by two contingency table. For a single binomial parameter, Jeffrey's prior is the Beta (0.5,0.5) distribution. For multinomial probabilities, Jeffrey's prior is the Dirichlet distribution with all parameters equal to 0.5. In the following sections, we briefly review standard Wald and adjusted Wald

intervals for association parameters of interest. We then describe Bayesian methods for credible interval construction with Jeffreys priors as well as with proposed alternative Dirichlet priors that are designed to improve coverage. Simulations were performed to compare the performance of each method, with a focus on interval width and frequentist coverage. The goal of the study is to find out if Bayesian methods with Jeffreys priors and/or alternative Dirichlet priors show improvements compared to standard methods.

2.Methods

We consider a cross-sectional sample producing a 2*2 table for associating a binary outcome with a binary exposure:

		Exposure	
		Yes	No
Outcome	Yes	N_{11}	N_{10}
	No	N_{01}	N_{00}

		Exposure	
		Yes	No
Outcome	Yes	P_{11}	P_{10}
	No	P_{01}	P_{00}

Above, the table on the left reflects the multinomial cell counts and the table on the right provides a notation for the true cell probabilities. Then we have:

$$\text{Risk Difference (RD)} = \frac{P_{11}}{P_{11}+P_{01}} - \frac{P_{10}}{P_{10}+P_{00}} \quad (1)$$

$$\text{Odds Ratio (OR)} = \frac{P_{11}P_{00}}{P_{01}P_{10}} \quad (2)$$

$$\text{Relative Risk (RR)} = \frac{P_{11}(P_{10}+P_{00})}{P_{10}(P_{11}+P_{01})} \quad (3)$$

The MLEs for the three parameters are obtained by replacing the P_{ij}s by the corresponding cell counts (n_{ij}s) that are obtained in the sample.

2.1 Standard Wald intervals

The most commonly used method to calculate the confidence intervals is inverting large sample Wald tests. It evaluates the standard errors at the maximum likelihood estimates. By using delta method [11] and assuming a natural four-cell multinomial model for the cell counts in the two-by-two table, the three estimators are approximately normal with asymptotic standard errors.

For risk difference, $SE\{\widehat{RD}\} = \sqrt{\widehat{Var}(\widehat{p_1}) + \widehat{Var}(\widehat{p_2}) - 2Cov(\widehat{p_1}, \widehat{p_2})}$. Since the MLEs for p_1 and p_2 are uncorrelated binomial proportions,

$$SE\{\widehat{RD}\} = \sqrt{\widehat{Var}(\widehat{p_1}) + \widehat{Var}(\widehat{p_2})} = \sqrt{\frac{p_1(1-p_1)}{n_{11}+n_{01}} + \frac{p_2(1-p_2)}{n_{10}+n_{00}}} = \sqrt{\frac{n_{11}n_{10}}{(n_{11}+n_{10})^3} + \frac{n_{10}n_{00}}{(n_{10}+n_{00})^3}} \quad (4)$$

For the odds ratio and relative risk, one can apply the multivariate delta method:

$$Var\big(f(x)\big) = \widehat{D}'\widehat{Var}([n_{11}, n_{10}, n_{01}, n_{00}])\widehat{D}$$

$$\widehat{D} = [\frac{\widehat{df}}{dn_{11}}, \frac{\widehat{df}}{dn_{10}}, \frac{\widehat{df}}{dn_{01}}, \frac{\widehat{df}}{dn_{00}}]$$

$$f(x_1) = \log(\widehat{OR}) = \log(n_{11}) + \log(n_{00}) - \log(n_{10}) - \log(n_{01})$$

$$f(x_2) = \log(\widehat{RR}) = \log(n_{11}) + \log(n_{00} + n_{10}) - \log(n_{10}) - \log(n_{01} + n_{11})$$

After algebra, we have:

$$\text{SE}\{\ln(\widehat{OR})\} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{01}} + \frac{1}{n_{10}} + \frac{1}{n_{00}}} \quad (5)$$

$$\text{SE}\{\ln(\widehat{RR})\} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{10}} - \frac{1}{n_{11}+n_{01}} - \frac{1}{n_{10}+n_{00}}} \quad (6)$$

The approach of using the delta method-based SEs to calculate confidence intervals for the odds ratio and relative risk is known as Woolf's method [12]. The corresponding approximate 95% confidence intervals are:

95% CI for RD = RD – 1.96*SE{RD} to RD + 1.96*SE{RD} (7)

95% CI for OR = exp(ln(OR) – 1.96*SE{ln(OR)}) to exp(ln(OR) + 1.96*SE{ln(OR)}) (8)

95% CI for RR = exp(ln(RR) – 1.96*SE{ln(RR)}) to exp(ln(RR) + 1.96*SE{ln(RR)}) (9)

2.2 Gart intervals and adjusted Gart intervals

Wald intervals often behave poorly for small samples, exhibiting coverage probabilities that are too low. They can remain deficient even if the sample is relatively large, especially in categorical settings involving relatively small or large probabilities. Gart provided one method which works better for small samples [7]. This method adds 0.5 to each cell when conducting both the point and standard error estimation. Thus the problems with computation of effects or standard errors caused by zero cells are solved, and the coverage probabilities are improved. In

this project, we also want to examine the alternative of only adding 0.5 to 0 cells in the point and standard error estimation (leaving non-zero cell counts as they are). We will call this method “adjusted Gart”. These two methods are included in our simulation comparison.

2.3 Bayesian credible intervals

Agresti compared and summarized some commonly used priors for proportions [9]. As he said, even though a relatively informative prior can represent the researcher’s subjective beliefs, it may cause poor performance in terms of ordinary frequentist criteria especially when the prior beliefs are incorrect. He recommended using quite diffuse priors in order to maintain good coverage performance, and suggested that even uniform priors are too informative. Lyles, Weiss and Waller (2020) examined credible intervals based on the two most popular Bayesian priors for p. These are the weakly informative uniform and Jeffreys priors. Since both of them lead to beta posteriors, sampling is not needed and credible intervals can be calculated using any software program that provides access to beta distribution percentiles. Among the two options, the Jeffreys prior is highly recommended for satisfactory average coverage properties (e.g., Brown et al. 2001). It is generally seen as performing well even for high and low p. Lyles et al. (2020) affirmed the notion that the intervals based on the uniform and Jeffreys prior obtain favorable overall coverage on average across the full range of p.

However, they uncovered notable deficiencies in terms of coverage balance. For example, for $p < 0.5$, Jeffreys prior reduces high-side errors characteristic of the uniform prior, but at the expense of low-side errors. In order to achieve specified nominal coverage criteria for both high-side and low-side errors, Lyles et al. (2020) provided an alternative which selects an optimal value κ between (0,0.5) and uses Beta(κ, $1-\kappa$) and Beta($1-\kappa$, κ) priors to calculate the lower bound and upper bound of the credible interval. In this study, we use a variant on this approach somewhat analogous to choosing $\kappa=0.25$, to produce credible intervals for association parameters that will be more conservative than those based on Jeffreys prior and yet less conservative than the Gart and adjusted Gart intervals.

As we know, the cell counts in the contingency table based on cross-sectional sampling are reasonably assumed to follow the multinomial distribution:

$$(N_{11}, N_{10}, N_{01}, N_{00}) \sim \text{Multinomial}\ (N, p_{11}, p_{10}, p_{01}, p_{00}) \qquad (10)$$

If we decide to place a Jeffreys prior on $\mathbf{p} = (p_{11}, p_{10}, p_{01}, p_{00})$, this implies the following assumption:

$$(p_{11}, p_{10}, p_{01}, p_{00}) \sim \text{Dirichlet}\ (0.5, 0.5, 0.5, 0.5) \qquad (11)$$

The posterior distribution is then known also to follow a Dirichlet distribution, i.e.,

$$p \mid n \sim \text{Dirichlet}\ (n_{11} + 0.5, n_{10} + 0.5, n_{01} + 0.5, n_{00} + 0.5) \qquad (12)$$

In addition to considering the standard Jeffreys prior, we adjust the Dirichlet parameters somewhat analogously to the approach taken by Lyles et al. (2020), in

order to make the credible intervals more conservative. The two sides of the confidence interval were obtained from two Dirichlet distributions.

For the upper bound, the prior on $(p_{11}, p_{10}, p_{01}, p_{00})$ is

$$(p_{11}, p_{10}, p_{01}, p_{00}) \sim \text{Dirichlet}(0.75, 0.25, 0.25, 0.75) \quad (13)$$

Thus the posterior distribution is

$$p|n \sim \text{Dirichlet}(n_{11} + 0.75, n_{10} + 0.25, n_{01} + 0.25, n_{00} + 0.75) \quad (14)$$

The 97.5^{th} percentile of the posterior distribution was used as the upper limit.

For the lower bound, the prior on $(p_{11}, p_{10}, p_{01}, p_{00})$ is

$$(p_{11}, p_{10}, p_{01}, p_{00}) \sim \text{Dirichlet}(0.25, 0.75, 0.75, 0.25) \quad (15)$$

Thus the posterior distribution is

$$p|n \sim \text{Dirichlet}(n_{11} + 0.25, n_{10} + 0.75, n_{01} + 0.75, n_{00} + 0.25) \quad (16)$$

The 2.5^{th} percentile of the posterior distribution was used as the lower limit.

We expected improvement after conducting interval estimation base on the adjusted Jeffreys priors, as they are designed to confer a measure of conservativeness.

2.4 Simulation process

The SAS statistical package V9.4 was used for generating the simulations. Wald and Gart intervals were simulated within the SAS IML procedure. A SAS macro generating Dirichlet random variables was built. Credible intervals were estimated by using the 2.5th and 97.5th percentiles of the posterior distributions as the two

sides of the intervals. 30000 simulations were generated in the study. The sum of frequencies in the contingency table was set to be 20, in order to simulate small sample situations.

During each simulation, 4 random multinomial probabilities were created. The mean of each randomly generated probability is 0.25, but these true probabilities were generated in such a way that they could each vary over the (0,1) range. Based on the multinomial distribution with each such set of probabilities, the numbers in the cells were generated randomly. Then the Wald intervals can be estimated by (7), (8) and (9), along with the Gart intervals (adding 0.5 to each cell prior to calculations) and adjusted Gart intervals (only adding 0.5 to each zero cell prior to calculations). Simulation runs where there was at least one cell with a 0 count are excluded when estimating and evaluating standard Wald intervals based on Woolf's method (as in Agresti and Min, 2005). For credible intervals, draws from the appropriate Dirichlet distributions were randomly generated based on the cell counts, (12) and (14). We set any posterior cell count that was less than 0.5 equal to 0.5. Then the intervals were constructed as described above, based on percentiles of the Dirichlet posterior distributions.

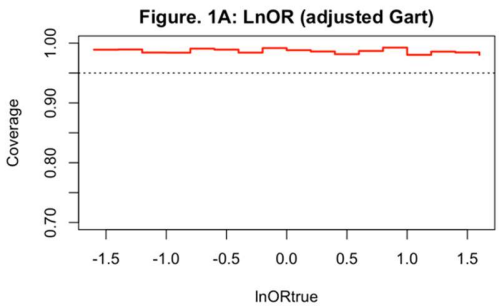
3. Results

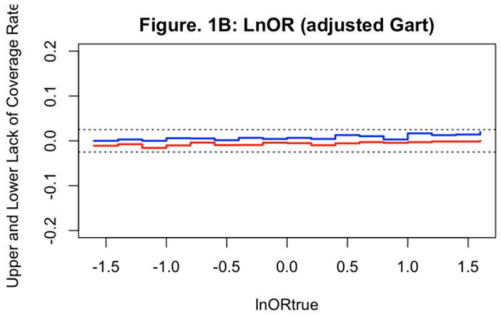
To compare the performance of the intervals, we initially selected the simulations where the generated true ln(OR) was between -1.6 and 1.6 (odds ratio from 0.2 to 5.0) and divided this range into 16 intervals with a step of 0.2. Since most epidemiology problems in real studies yield an odds ratio in this interval, it makes this study representative. Average coverage probabilities were calculated within each small interval, along with upper and lower lack of coverage rates and mean interval widths. A similar process was conducted for simulations where the true ln(OR) was between -3 and 3 (odds ratio from 0.05 to 20), as we also wish to consider the relative performance of the competing methods when odds ratios can be extremely high or low.

Figure 1 plots the overall coverage rates and the upper and lower lack of coverage vs lnORtrue for the 95% Gart, adjusted Gart, Cred and adjusted Cred intervals for ln(OR), when n=20. The Gart and adjusted Gart approaches lead to similar results. Their overall coverage rates are a lot higher than 95% (panel A and C), and their region-specific average upper and lower lack of coverage rates are always less than 0.025 (panel B and D). In contrast, the standard Bayesian intervals based on Jeffreys Dirichlet prior exhibit low coverage rates (panel E). Note that the upper lack of coverage rate increases while the lower lack of coverage rate decreases

when the absolute value of lnORtrue increases (panel F). The Bayesian intervals based on our proposed adjusted Jeffreys prior performed much better. Their region-specific average coverage rates are quite close to 0.95 (panel G). The upper and lower lack of coverage rates remain much closer to the dotted lines (panel H) than those for the standard credible intervals.

Comparisons of mean widths of confidence intervals were also conducted. The results indicated that the two Gart intervals are far more conservative than the two credible intervals (panel I and J). Among the two credible intervals, the one based on the standard Jeffreys prior is less conservative in terms of width, but unacceptably anticonservative in terms of coverage. The width advantage of the adjusted credible interval compared to the two Gart intervals is pronounced.





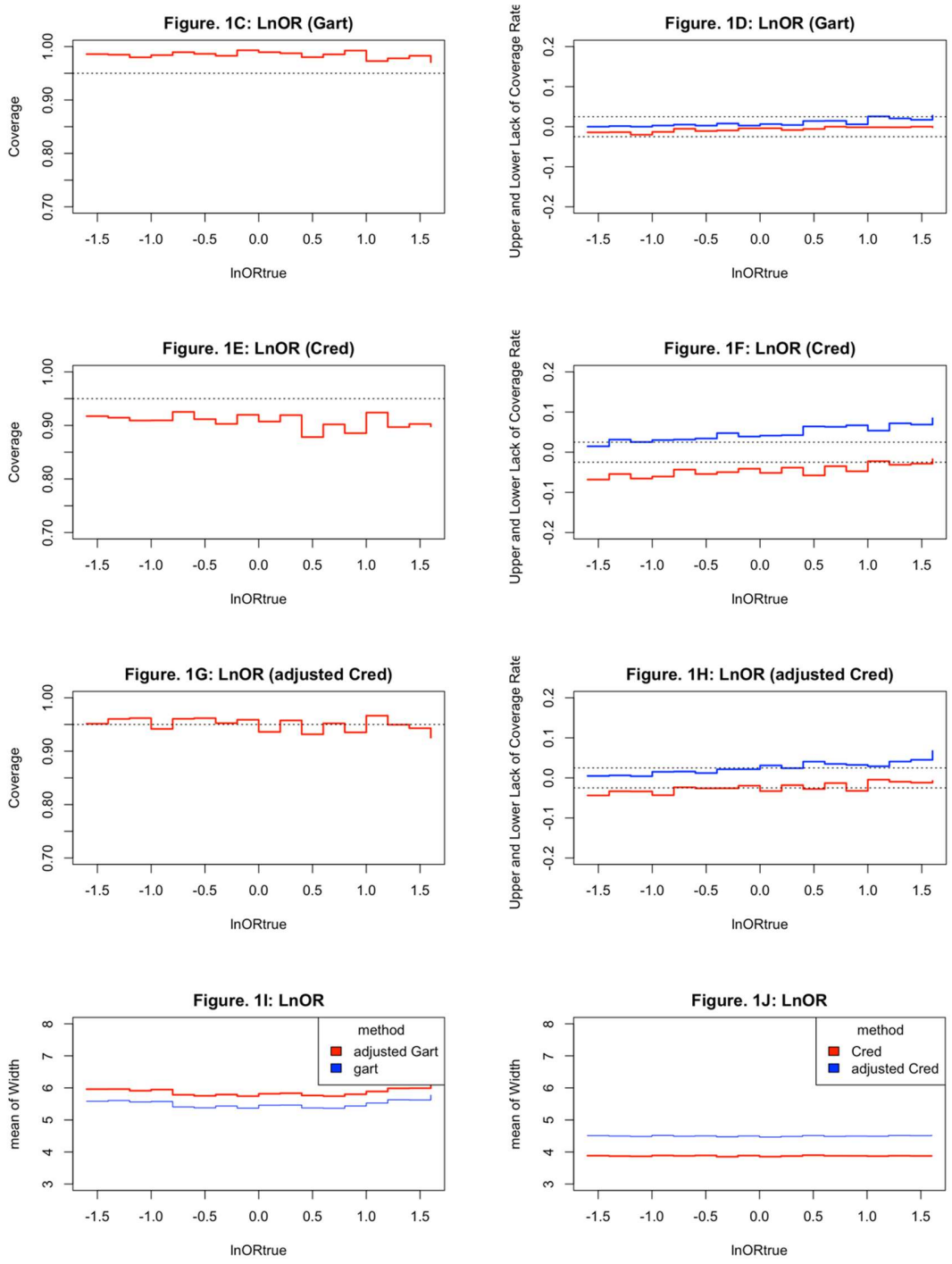


Figure 1 Overall coverage rates of 95% adjusted Gart (panel A), standard Gart (panel C), credible (panel E) and adjusted credible (panel G) intervals plotted over the range of LnORtrue from -1.6 to 1.6 for n=20, together with upper and lower lack of coverage rates for these intervals (panels B, D, F and H) and mean of width of the intervals (panel I and J). Positive y-axis values in panels B, D, F, H represent upper excursion probabilities and negative y-axis values represent lower excursion

probabilities (e.g, a value at 0.05 means the intervals misses high 5% of the time at that value of p; a value at -0.05 means the interval misses low 5% of the time at that value of LnORtrue). Dashed lines are drawn at 0.95 (panel A, C, E and G) and ∓ 0.025 (panel B, D, F and H).

Figure 2 plots the overall coverage rates and the lack of coverage vs lnORtrue for the 95% Gart, adjusted Gart, Cred and adjusted Cred intervals of risk difference (RD), when n=20. Again, the Gart and adjusted Gart intervals are close. Their overall coverage rates are consistently a little higher than 95% (panel A and C) and their upper and lower lack of coverage rate are almost always less than 0.025 (panel B and D). The Bayesian intervals with Jeffreys prior again produce low coverage rates (panel E). Both the upper and lower lack of coverage rates are consistently out of the desired range, which is -0.025 to 0.025 (panel F). The Bayesian intervals with the proposed adjusted Jeffreys prior performed well. Their overall coverage rates are quite close to 0.95 (panel G), and the lack of coverage rates are close to the dotted lines (panel H). Comparisons of the mean widths of the intervals again demostrate that the two Gart intervals are more conservative than the two credible intervals. Among the two credible intervals, the one based on the standard Jefferys prior is again the narrowest but is far too anti-conservative in terms of coverage.

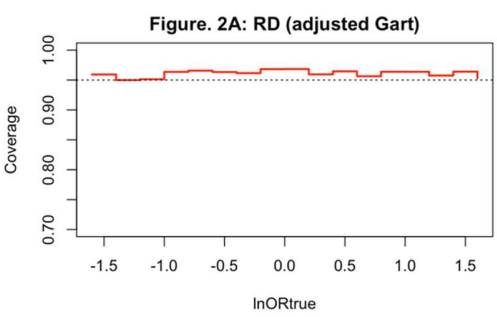


Figure. 2A: RD (adjusted Gart)
Coverage
0.70
0.80
0.90
1.00
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
lnORtrue

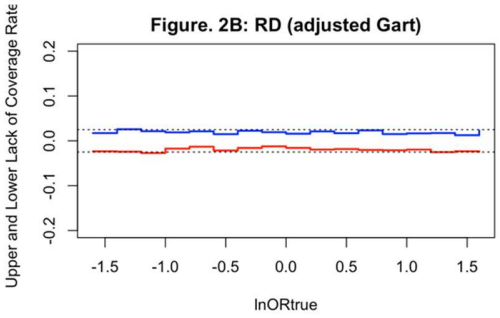


Figure. 2B: RD (adjusted Gart)
Upper and Lower Lack of Coverage Rate
-0.2
-0.1
0.0
0.1
0.2
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
lnORtrue

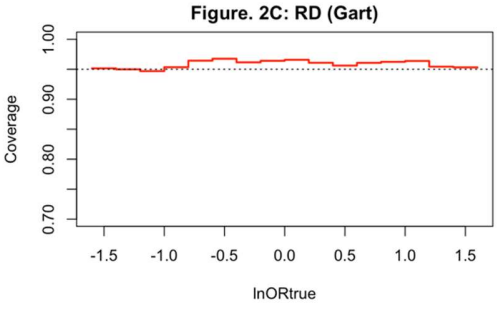


Figure. 2C: RD (Gart)
Coverage
0.70
0.80
0.90
1.00
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
lnORtrue

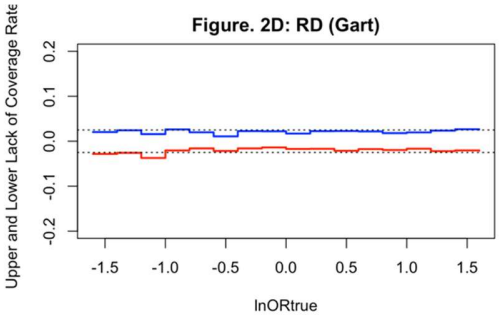


Figure. 2D: RD (Gart)
Upper and Lower Lack of Coverage Rate
-0.2
-0.1
0.0
0.1
0.2
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
lnORtrue

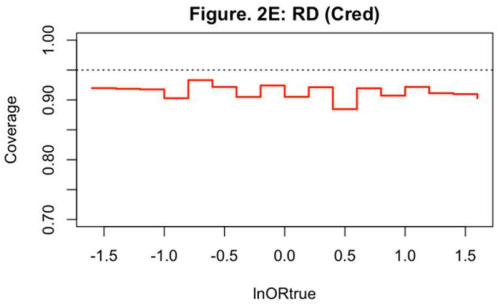


Figure. 2E: RD (Cred)
Coverage
0.70
0.80
0.90
1.00
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
lnORtrue

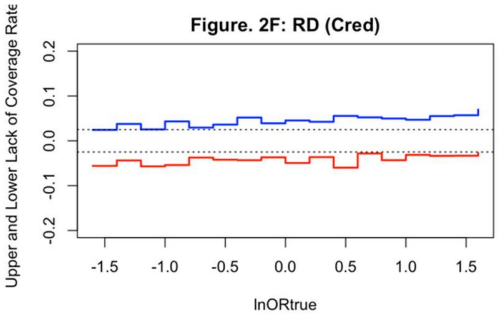


Figure. 2F: RD (Cred)
Upper and Lower Lack of Coverage Rate
-0.2
-0.1
0.0
0.1
0.2
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
lnORtrue

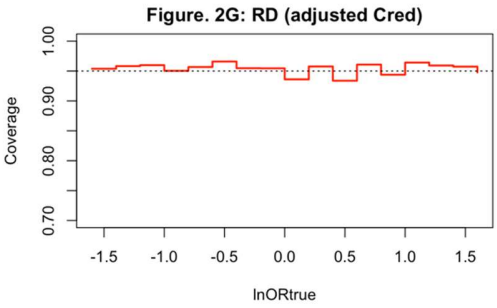


Figure. 2G: RD (adjusted Cred)
Coverage
0.70
0.80
0.90
1.00
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
lnORtrue

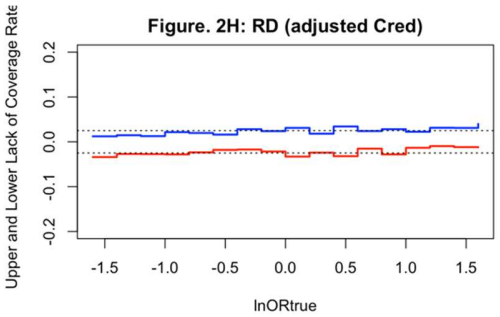


Figure. 2H: RD (adjusted Cred)
Upper and Lower Lack of Coverage Rate
-0.2
-0.1
0.0
0.1
0.2
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
lnORtrue

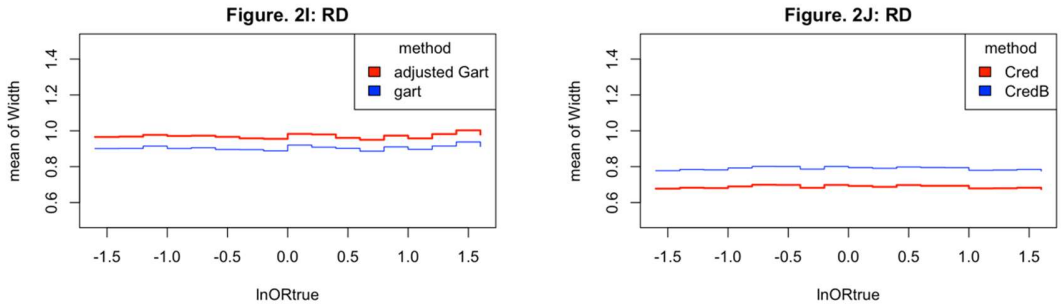
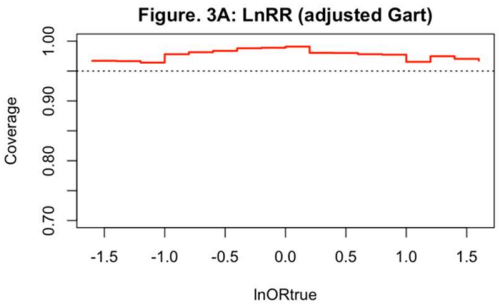
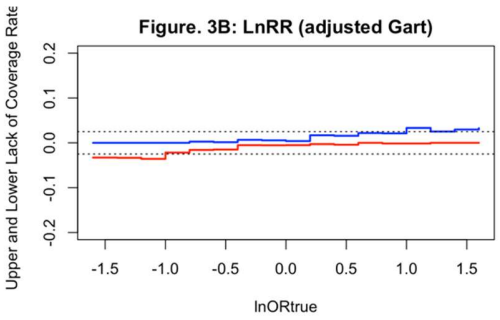


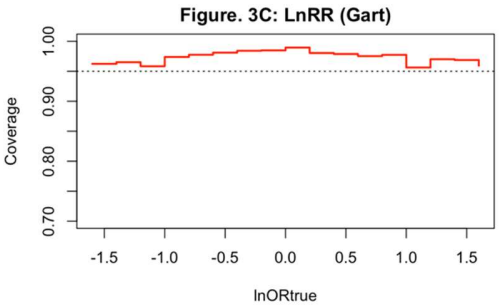
Figure 2 Overall coverage rates of 95% adjusted Gart (panel A), standard Gart (panel C), credible (panel E) and adjusted credible (panel G) intervals plotted over the range of LnORtrue from -1.6 to 1.6 for n=20, together with upper and lower lack of coverage rates for these intervals (panels B, D, F and H) and mean of width of the intervals (panel I and J). Positive y-axis values in panels B, D, F, H represent upper excursion probabilities and negative y-axis values represent lower excursion probabilities (e.g, a value at 0.05 means the intervals misses high 5% of the time at that value of p; a value at -0.05 means the interval misses low 5% of the time at that value of LnORtrue). Dashed lines are drawn at 0.95 (panel A, C, E and G) and ∓ 0.025 (panel B, D ,F and H).

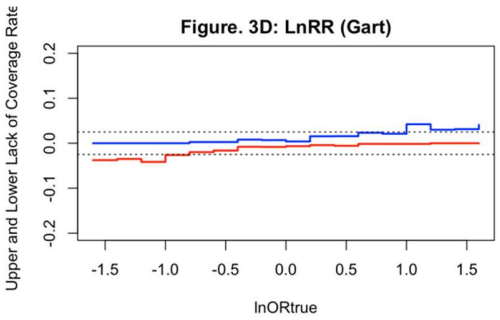
Figure 3 plots the overall coverage rates and the lack of coverage vs lnORtrue for the 95% Gart, adjusted Gart, Cred and adjusted Cred intervals of relative risk(RR) when n=20. The Gart and adjusted Gart intervals are again quite close. The Bayesian intervals with Jeffreys prior demonstrate low overall coverage rates (panel E), and both the upper and lower lack of coverage rates are out of the desired -0.025 to 0.025 range (panel F). The Bayesian intervals with adjusted Jeffreys prior again performed much better. Their coverage rates are quite close to 0.95 (panel G) overall, and the upper and lower lack of coverage rates remain close to the dotted line (panel H). Comparisons of mean width of confidence intervals again show that the two Gart intervals tend to be far wider than the two credible intervals. Among the two credible intervals, the one using the standard

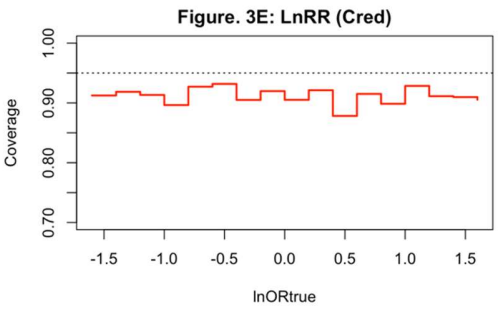
Jeffreys prior is less conservative and associated with overly narrow intervals.

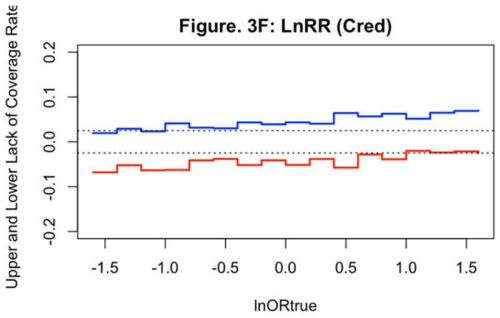


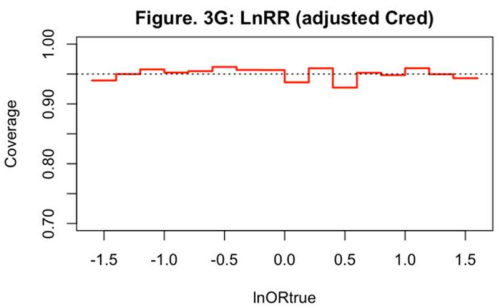


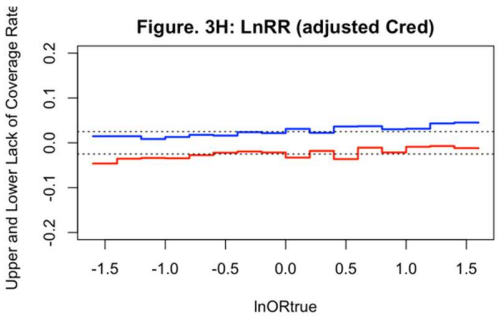












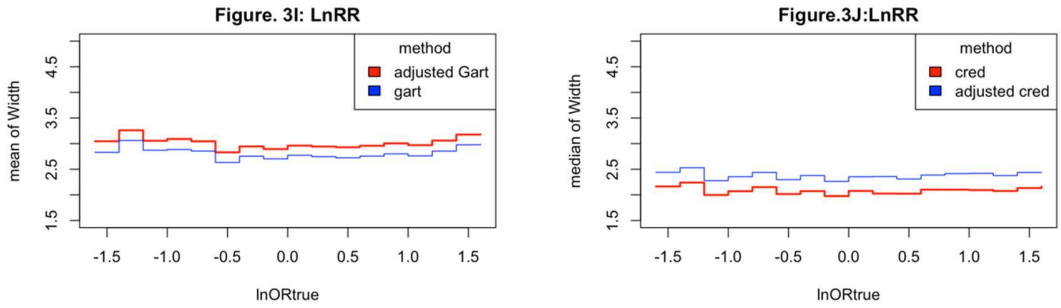
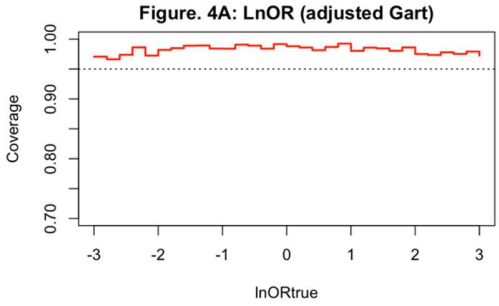
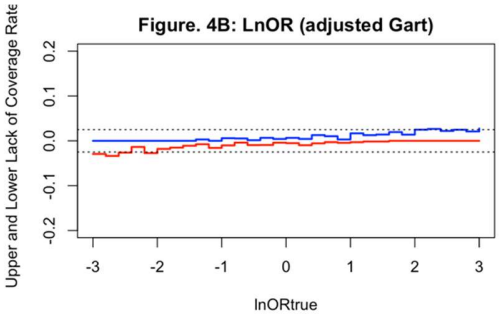


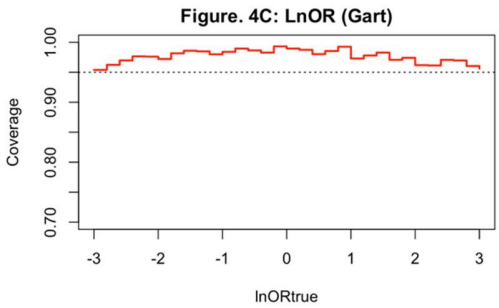
Figure 3 Overall coverage rates of 95% adjusted Gart (panel A), standard Gart (panel C), credible (panel E) and adjusted credible (panel G) intervals plotted over the range of LnORtrue from -1.6 to 1.6 for n=20, together with upper and lower lack of coverage rates for these intervals (panels B, D, F and H) and mean of width of the intervals (panel I and J). Positive y-axis values in panels B, D, F, H represent upper excursion probabilities and negative y-axis values represent lower excursion probabilities (e.g, a value at 0.05 means the intervals misses high 5% of the time at that value of p; a value at -0.05 means the interval misses low 5% of the time at that value of LnORtrue). Dashed lines are drawn at 0.95 (panel A, C, E and G) and ∓ 0.025 (panel B, D, F and H).

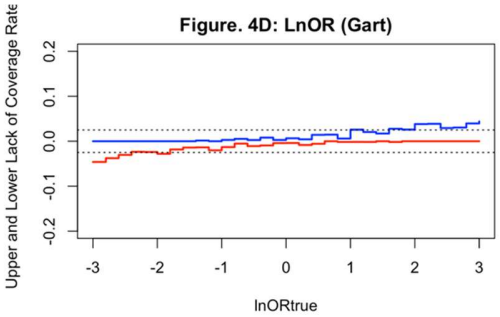
Figure 4 plots the overall coverage rates and the upper and lower lack of coverage vs lnORtrue for the 95% confidence intervals of LnOR, for all generated values of lnORtrue that fell between the wider range of -3 and 3. We increased the range of lnORtrue to find out how these intervals perform when the odds ratio is extremely small or large. When the absolute value of the lnOR increases, the overall coverage rates for the Gart intervals continue to display conservativeness. However, we note that the credible intervals show more of a trend toward anti-conservativeness as the OR becomes quite extreme. For all intervals, we note that one of the upper or lower lack of coverage rates trends toward being relatively high while the other one is close to 0 if you compare these Figure 4 to Figure 1. The means widths of the Gart and adjusted Gart methods increase when the absolute

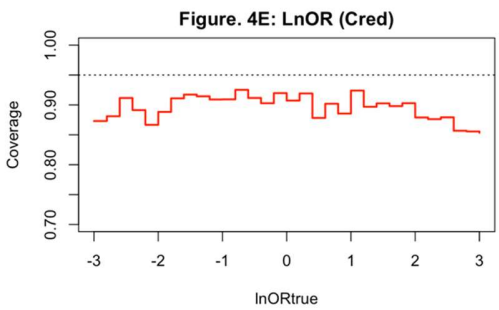
value of lnOR increases, while the mean widths of the two credible intervals remain almost unchanged. This makes logical sense given the decreasing coverage trends in the extreme in panels 4E and 4G.

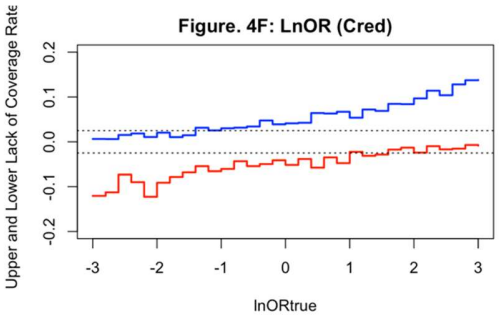












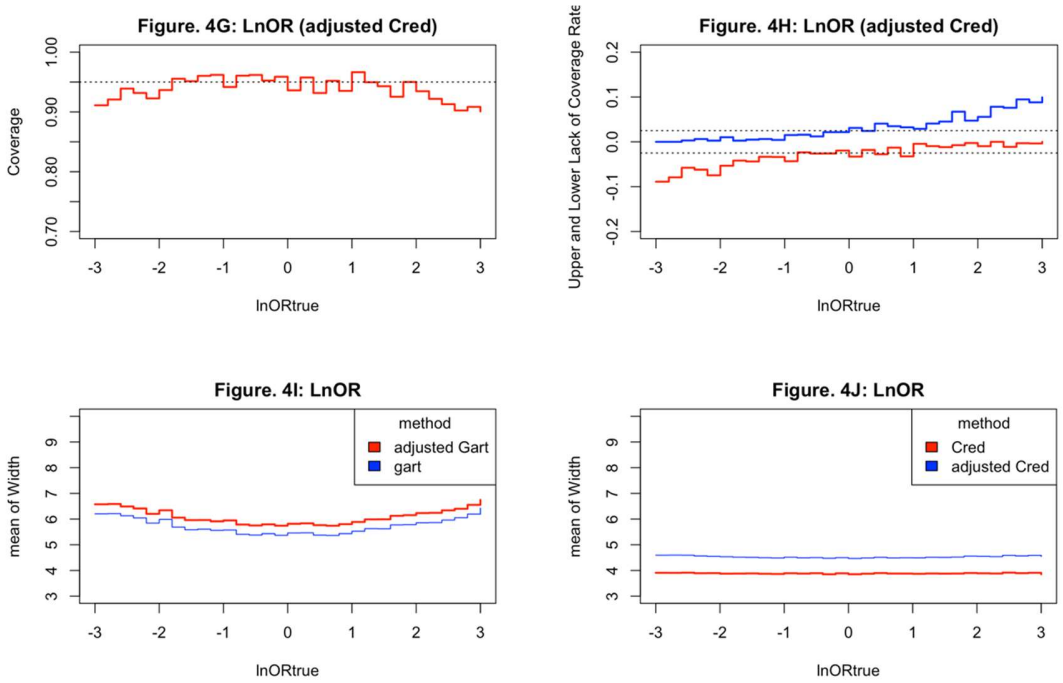


Figure 4 Overall coverage rates of 95% adjusted Gart (panel A), standard Gart (panel C), credible (panel E) and adjusted credible (panel G) intervals plotted over the range of LnORtrue from -3 to 3 for n=20, together with upper and lower lack of coverage rates for these intervals (panels B, D, F and H) and mean of width of the intervals (panel I and J). Positive y-axis values in panels B, D, F, H represent upper excursion probabilities and negative y-axis values represent lower excursion probabilities (e.g, a value at 0.05 means the intervals misses high 5% of the time at that value of p; a value at -0.05 means the interval misses low 5% of the time at that value of LnORtrue). Dashed lines are drawn at 0.95 (panel A, C, E and G) and ∓ 0.025 (panel B, D, F and H).

4. Discussion

Although there are some drawbacks, Wald-type intervals are commonly used and offer a simple way to find confidence intervals for OR, RR and RD based on cross-sectional data associating a binary health-related outcome with a binary exposure. Interestingly, we found that both the Gart and adjusted Gart intervals were conservative across the board over a wide range of true risk parameters when

sample size was small (N=20).

Our study found, somewhat surprisingly, that a standard Bayesian credible interval based on the Jeffreys Dirichlet prior produces highly anti-conservative results over the same range of parameters. Our proposed adjusted credible interval provides a far more favorable alternative that mitigates the conservative lack of coverage problems of the Wald method in some cases where sample size is small and/or the population is characterized by relatively small or large probabilities. In general, we have to make a trade between coverage rates and the width of the confidence interval. For N=20, our proposed adjusted Bayesian approach gives credible intervals with generally favorable overall coverage rates that are far less conservative (i.e., markedly narrower) than the Gart intervals. It generally performed well over the range between 0.2 and 5 for the true odds ratio.

Previous studies used Bayesian methods based on some priors such as Jeffreys priors, logit-normal priors, and related correlated priors to compare proportions in two-by-two contingency tables[9,13]. However, to our knowledge this is the first attempt to consider adjusting the beta priors in a tailored way to control both the lower and upper coverage rates and apply this method to OR, RR and RD. Although this approach is a little more conservative than the one based on Jeffreys prior, it reduces the lack of coverage on both the upper and lower sides greatly.

As for limitations, Figure 4 reveal a clear tendency to miss on the high side when the true OR gets large and a corresponding tendency to miss on the low side for small true OR values. This problem exists for all of the methods we used in the study. It indicates a lack of coverage balance when the odds ratio is too high or too small. Also, it appears that interval width can be reduced markly at a relatively little risk of the coverage probability falling much below the nominal confidence level by using Bayesian methods, if adjustments like those proposed here are made. Nevertheless, if the analyst can not tolerate any possibility that the coverage rate falls below the nominal confidence level, then the Gart and adjusted Gart methods would be better. Our results suggest that the standard Gart method could be preferred over the adjusted version, since it produces narrower intervals with coverage that remains conservative.

For future work, we are interested to see how well the proposed adjusted credible interval works compared to Wald methods over a broader range of sample sizes. Also, we want to focus on variations on the proposed Dirichlet priors and to investigate if there are better priors which improve the coverage propertites while maintating interval width benefits. Some sort of optimization might also be considered (e.g., Lyles et al. 2020), if an efficient approach that avoids direct sampling from Dirichlet posteriors can be incorporated.

Bibliography

1. Levin, K. A. Study design III: Cross-sectional studies. *Evid. Based. Dent.* **7**, 24–25 (2006).
2. Kodde, D. A. & Palm, F. C. Wald criteria for jointly testing equality and inequality restrictions. *Econom. J. Econom. Soc.* 1243–1248 (1986).
3. Ghosh, B. K. A comparison of some approximate confidence intervals for the binomial parameter. *J. Am. Stat. Assoc.* **74**, 894–900 (1979).
4. Vollset, S. E. Confidence intervals for a binomial proportion. *Stat. Med.* **12**, 809–824 (1993).
5. Brown, L. D., Cai, T. T. & DasGupta, A. *Confidence intervals for a binomial proportion and Edgeworth expansions*. (1999).
6. Agresti, A. & Caffo, B. Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *Am. Stat.* **54**, 280–288 (2000).
7. Gart, J. J. & Nam, J. Approximate interval estimation of the ratio of binomial parameters: a review and corrections for skewness. *Biometrics* 323–338 (1988).
8. Carlin, B. P. & Louis, T. A. *Bayes and empirical Bayes methods for data analysis*. (Chapman and Hall/CRC, 2010).
9. Agresti, A. & Min, Y. Frequentist performance of Bayesian confidence

intervals for comparing proportions in 2× 2 contingency tables. *Biometrics* **61**, 515–523 (2005).

10. Brown, L. D., Cai, T. T. & DasGupta, A. Interval estimation for a binomial proportion. *Stat. Sci.* 101–117 (2001).

11. Cox, C. Delta method. *Encycl. Biostat.* **2**, (2005).

12. Woolf, B. On estimating the relation between blood group and disease. *Ann Hum Genet* **19**, 251–253 (1955).

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