ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

DOI: 10.5336/biostatic.2016-52932

An Empirical Bayesian Approach in Estimating Odds Ratios for Rare or Zero Events

Nadir ya da Sıfır Olaylar İçin Göreli Oranların Tahmininde Ampirik Bayesci Bir Yaklaşım

Mehmet KOÇAK[a]

[a]Department of Preventive Medicine, Division of Biostatistics, The University of Tennessee Health Science Center, Memphis, USA

Geliş Tarihi/*Received:* 06.09.2016
Kabul Tarihi/*Accepted:* 14.12.2016

Yazışma Adresi/*Correspondence:* Mehmet KOÇAK
The University of Tennessee Health Science Center, Department of Preventive Medicine, Division of Biostatistics, Memphis, USA/ABD
mkocak1@uthsc.edu

ABSTRACT Objective: For rare events, one or two cells of 2x2 tables may have zero (0) counts or a count very close to zero. This issue makes the estimation of Odds Ratio (OR) impossible or leads to unstable OR estimates due to complete or quasi-complete separation. In such cases, Exact Logistic Regression and Firth approach may reduce the bias but these two methods do not resolve the issue completely. **Material and Methods:** In this research, we propose an Empirical Bayesian estimation procedure for Odds Ratio of rare events and provide a formal test procedure. We compare our Bayesian approach with Exact Logistic and Firth approaches in terms of statistical power achieved through extensive simulations using no event and varying event rate scenarios. **Results:** We show that our Bayesian approach retains the Type-I error rate and is much more powerful than the currently existing methods such as Exact Logistic Method and Firth approach. We also show that an Odds Ratio estimate is possible with our Bayesian approach even in no event or extremely rare events cases where the other two approaches suffer. **Conclusion:** Our new Bayesian estimation method is more powerful than Exact Logistic Regression and Firth approaches and provides an Odds ratio estimate even in no event or extremely rare events cases.

Keywords: Bayesian Odds ratio; zero cell count; rare events; Bayesian approach; statistical simulation; exact logistic regression

ÖZET Amaç: Nadir olaylar için 2x2'lik tablolarda bir ya da iki hücre sıfır (0) ya da sıfıra çok yakın olabilir. Bu sorun, Göreli Oranın tahmininin imkansızlığına ya da tam veya yarı-ayrılma nedeniyle güvenilir olmayan Göreli Oran tahminlerine neden olur. Bu gibi durumlarda, Kesin Lojistik Regresyon ve Firth yaklaşımları yanlılığı azaltabilir fakat bu iki yöntem de sorunu tamamen ortadan kaldırmaz. **Gereç ve Yöntemler:** Bu araştırmada, nadir olayların Göreli Oranı için ampririk bir Bayesci tahmin yöntemi ve yeni bir test prosedürü sunuyoruz. Bayesci yaklaşımımızı, Kesin Lojistik Regresyon ve Firth yaklaşımlarıyla, istatistiksel güç açısından, sıfır olay ve ileri derecede nadir olaylı senaryoların işlendiği geniş kapsamlı simülasyonlarla karşılaştırıyoruz. **Bulgular:** Bizim Bayesci yöntemimizin 1. Tip hatayı koruduğunu ve Kesin Lojistik Regresyon ve Firth yaklaşımlarına göre daha güçlü olduğunu gösterdik. Aynı zamanda, Bayesci yöntemimizin Göreli Oranını, diğer iki yöntemin sınıfta kaldığı, sıfır olay ve ileri derecede nadir olaylar durumlarında da tahmin edebildiğini gösterdik. **Sonuc:** Bizim yeni Bayesci tahmin yöntemimiz, Kesin Lojistik Regresyon ve Firth yaklaşımlarından istatistiksel olarak daha güçlüdür ve sıfır olay ile, ileri derecede nadir olaylar durumlarında bile, bir Göreli Oranı tahmini verebilir.

Anahtar Kelimeler: Göreli oran; sıfır hücre sayısı; nadir olaylar; Bayesci yaklaşım; istatistiksel simülasyon; kesin lojistik regrasyon

Like in many other application areas, in clinical trials and health sciences as well, Odds Ratio (OR) is one of the most commonly used statistical measure of association between an independent binary factor and a binary response.

Suppose we have two treatment arms in a protocol and we are interested in the association of Treatment with a binary response variable. We can represent this by a 2x2 table as in Table 1.

TABLE 1. A representative 2x2 table.

		Response		
		Yes	No	
Treatment	Arm-1	a	b	a+b
	Arm-2	c	d	c+d
		a+c	b+d	n=a+b+c+d

In such a setup, the probability of Response (i.e., Response=YES) is estimated by

$$\hat{p}_1 = \frac{a}{a+b}, \quad \hat{p}_2 = \frac{c}{c+d}.$$

From this, we estimate OR and logged version of OR as follows:

$$\overline{\text{OR}} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2(1-\hat{p}_2)}, \text{and} \log(\overline{\text{OR}}) = \overline{LOR} = \log(\frac{\hat{p}_1}{1-\hat{p}_1}) - \log(\frac{\hat{p}_2}{1-\hat{p}_2})$$

The variance estimate of the logged-OR can also be obtained as $V(\overline{\text{LOR}}) \cong \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$.

From the above formulae, it is clear that if any of the cell in the 2x2 table is zero (0), then, maximum likelihood estimates of Odds Ratio and its variance do not exist. When a predictor, like Treatment, predicts the outcome variable perfectly, this is called a 'Complete Separation' and in such cases, Odds Ratio estimate does not exist even a given software package still aims at providing an estimate, which would not be a reliable estimate. When the separation happens at a lesser degree, then such separations are called 'quasi-complete' separations and makes the estimation process of Odds Ratios unstable and the resulting estimates unreliable. More on the maximum likelihood estimate of Odds Ratio and the issue of complete and quasi-complete separation can be found in Albert et al. (1984), Harrell et al. (1985), So (1993), and Agresti (1996).[1-4]

Naturally, this issue is easy to observe in univariable analyses like in 2x2 tables; however, the issue still exists and may not be easily noticeable in multivariable analyses such as in fitting logistic regression models. In modeling for a binary response variable, a set of one or more variables may perfectly predict the outcome, leading to complete or quasi-complete separation issue. This issue has been discussed by many among whom we mention Albert et al. (1984), Lesaffre et al. (1989) and Zorn (2005).[1,5,6]

To address this issue, Firth proposed a penalized maximum likelihood approach by penalizing the score equation.[7] This method reduces the bias of the odds ratio estimates and provide more reliable standard error estimates. Another approach to address this issue is to use the Exact Logistic Regression approach, which also reduces the bias with a much more conservative standard error estimates. A third approach to deal with quasi-complete or complete separation situation is the Bayesian logistic regression modeling, which can be implemented in SAS under GENMOD procedure. In Section 2, using a sample data, we will illustrate each of these methods in SAS and compare the results from each method.

In this paper, we propose an empirical Bayesian approach to estimate the Odds ratio and its corresponding significance test. We show through extensive simulations that our Empirical Bayesian approach is much more powerful compared to the Exact method and Firth Method even for rare events in both treatment arms which would result in zero (0) event for both arms. In Section 2, we describe two meth-

ods dealing with the zero cell count issue, namely, the Firth, the exact method, and the Bayesian Logistic Regression approaches. In Section 3, we describe our Bayesian approach and its testing algorithm. Section 4 presents the simulation setup and results followed by Section 5, where we provide an application to a real data. We finish with discussion and conclusions in Section 6. All computations in this research were conducted on SAS® Version 9.4.

TABLE 2: Sample clinical trials data.

	Favorable Outcome	Unfavorable Outcome	All Patients)
Placebo	0	36	36
New Drug	6	30	36
All Patients	6	66	72

EXISTING METHODS TO DEAL WITH ZERO CELL COUNT ISSUE

Suppose we have the following clinical trial results (Table 2):

```
data new2x2;
do id=1 to 0; treatment=1; outcome=1; output; end;
do id=1 to 36; treatment=1; outcome=2; output; end;
do id=1 to 6; treatment=2; outcome=1; output; end;
do id=1 to 30; treatment=2; outcome=2; output; end;
run;
proc format; value tx 1='Placebo' 2='New Drug';
value outcome 1='Favorable Outcome' 2='Unfavorable Outcome'; run;
```

Due to the zero (0) favorable response in the Placebo arm, estimating odds ratio and its standard error is not possible.

We first illustrate each available method using this toy example.

FIRTH APPROACH IMPLEMENTATION

Firth penalized likelihood approach is used to produce finite and consistent estimates of Odds Ratios in Logistic Regression case, when the maximum likelihood estimates may not be exist or consistent due to complete or quasi-complete separation.

Here is how we can implement the Firth approach and obtain the resulting Odds Ratio estimates:

```
proc logistic data=new2x2; format outcome outcome.;
model outcome (ref='Unfavorable Outcome')=treatment/rl firth; run;
```

From the Firth approach, we obtain an Odds Ratio estimate of 15.56 (95% CI: 0.809, 299.309) with a p-value of 0.0689.

EXACT LOGISTIC REGRESSION APPROACH IMPLEMENTATION

Exact Logistic Regression utilizes the likelihood of the observed response with respect to all 2^n possible response vectors. It is used for trials with small sample size and when event size is small, which leads to cells with no observation or small number of observations in contingency tables. Even with 30 observa-

tions, the number of response vectors to be considered reaches beyond one billion, and therefore, the procedure may get computationally costly very quickly.

On the same data, we can implement the Exact Logistic Regression approach as follows:

```
proc logistic data=new2x2; format outcome outcome.;
model outcome (ref='Unfavorable Outcome')=treatment/rl;
exact treatment/estimate=both; run;
```

From the Firth approach, we obtain an Odds Ratio estimate of 9.334 (95% CI: 1.694, infinity) with a p-value of 0.0249.

BAYESIAN LOGISTIC REGRESSION IMPLEMENTATION

Finally, here is how the Bayesian Logistic Regression model can be implemented:

```
data myprior; input _type_ $ Intercept treatment;
datalines;
Var 1 0.5
Mean 0 1
; run;
proc genmod data=new2x2 order=data;
model outcome= treatment/dist=binomial link=logit;
bayes seed=34367 plots=all nbi=20000 nmc=100000
thin=10 coeffprior=normal(input=myprior); run;
```

Here, the Odds Ratio is estimated to be 1.8 (95% CI: 0.35, 9.02).

As we see, the odds ratio estimates, its standard error, and thus its significance differ so much among the three approaches we described.

A NEW BAYESIAN APPROACH TO ESTIMATE ODDS RATIO

Now, we like to introduce our empirical Bayesian approach.

Let p_1 be the success probability for Population-1 (say, 'placebo' arm) and p_2 the success probability for Population-2 (say, 'treatment' arm). We then assume the following prior distributions:

- $p_1 \sim Uniform(0,1)$
- $p_2 \sim Uniform(0,1)$

Let n_1 be a random sample from Population-1 with X_1 successes, and n_2 the success probability for Population-2 with X_2 successes. Assuming that the all observations satisfy the independence and identical distribution assumption, the likelihood function of (p_1, p_2) given such a data can be given as follows:

$$L(p_1, p_2 | X_1, X_2, n_1, n_2) = \binom{n_1}{X_1} p_1^{X_1} (1-p_1)^{n_1 - X_1} . \binom{n_2}{X_2} p_2^{X_2} (1-p_2)^{n_2 - X_2}$$

From the above prior selection and the likelihood function, we express the Posterior distribution of p_1, p_2 given the data as follows:

$$\prod(p_1, p_2 | X_1, X_2, n_1, n_2) \propto p_1^{X_1} (1-p_1)^{n_1 - X_1} .\ p_2^{X_2} (1-p_2)^{n_2 - X_2}$$

Given p_2, we see that $\prod(p_1|X_1, X_2, n_1, n_2, p_2) \propto p_1^{X_1}(1-p_1)^{n_1-X_1}$, which is a Beta-distribution with parameters (X_1+1, n_1-X_1+1). Similarly, given p_1, we see that $\prod(p_2|X_1, X_2, n_1, n_2, p_1) \propto p_2^{X_2}(1-$ *p2n2–X2*, which is a Beta-distribution with parameters *X2+1,n2–X2+1*. As these two distributions are independent-Beta distributions with means of $\frac{X_1+1}{(n_1-X_1+1)+(X_1+1)} = \frac{X_1+1}{n_1+2}$ and $\frac{X_2+1}{n_2+2}$, respectively.

Using the above results on the posterior distribution, we use the following steps to estimate the Odds Ratio and provide an Empirical testing procedure:

- Draw 1,000,000 samples (or more if desired) from $Beta(X_1+1, n_1-X_1+1)$ and $Beta(X_2+$ *1,n2–X2+1*, independently.
- Split these 1,000,000 samples into clusters of 1,000 samples.
- For each pair of (p_{1i}, p_{2i}) realization, compute the Log Odds Ratio (LOR) as

$$\log\left[\left(\frac{p_{1i}}{1-p_{1i}}\right) / \left(\frac{p_{2i}}{1-p_{2i}}\right)\right]$$

- Within each cluster, compute the median log Odds Ratio and the number of samples with the 2.5th percentile of LOR greater than zero (0), or the 97.5th percentile of LOR less than zero (0);
- Compute the median of the Median Log-Odds Ratios, the 2.5th percentile and 97.5th percentile, as well as the proportion of the significant cases in these 1,000 clusters, which will serve as the magnitude of significance (equivalent to p-value in frequentist sense);

All these statistics are computed and saved in a dataset to be used outside the Macro program; Here is a quick illustration of our SAS macro on the toy example we used earlier:

%***orbayes***(n11=**6**, n1=**30**, n21=**0**, n2=**36**, postN=10000000, gsize=10000, outname=pvalue);

proc print data=pvalue; **run**;

In the SAS Macro, the parameter 'postN' represents the number of samples drawn from the Beta-distribution components of the posterior distribution, and the parameter 'gsize' is the cluster size.

From the above analysis, we obtain the Median Odds Ratio estimate of 14.43 (95% CI: 1.99, 424.52) with a p-value <0.001.

Now, we like to compare our empirical Bayesian approach with the other three methods.

SIMULATION SETUP AND RESULTS

We have conduced our simulations with the following setup:

- ***P***1**(Probability of Success in Population-1):** 0.01 to 0.05 by 0.01
- **ODDS RATIO:** 1, 5, 9 (***P***2 is computed from the P1 and OR pair)
- **GROUP SAMPLE SIZE (Balanced group size):** N1=N2=50 to 500 by 50
- **GROUP SAMPLE SIZE (Unbalanced group size):** N1=500, N2=50 to 450 by 50
- Total of 290 combinations
- 1,000 random samples generated under each scenarios
- 1,000,000 posterior samples are generated for each simulation run.

For each simulation sample, we computed and recorded the OR estimate, its standard error, the resulting *empirical p-value*, and the number of times the estimation was possible.

RESULTS

We present the Type-1 error rate, empirical power, and Odds Ratios estimates from each model in Figures 1-5. The first observation we make is that the GENMOD approach has a positive bias in estimating the null Odds Ratio (OR=1), especially in extreme rare events scenarios with small sample size, which improves as the sample size and/or event rate increase. This issue of overestimation in the null case results in inflated Type-1 error rates (Figure 3) for the GENMOD approach. The Exact Logistic Regression approach also suffers from inflated Type-I error. Our Bayesian approach and the Firth approach estimate the OR close to 1.0, and thus, retain the Type-I error at or below 0.05.

All four methods we compared underestimate OR, which improves, as expected, when the event rate and sample size go up. Surprisingly, again, the GENMOD approach consistently underestimate OR compared to the other methods even with large sample sizes. Due to the inflated Type-1 error in the GENMOD approach, which would lead to artificial power increase naturally, we compare the empirical power only among the three remaining approaches, namely, Empirical Bayesian, Firth, and Exact Logistic Regression approaches.

Empirical power from the Bayesian approach is much superior to the Firth approach for small sample size cases, approaching to each other as sample size go up (Figure 4). Although the exact Logistic Regression approach seem to have reasonably competitive power for small sample sizes, it is inflated due to inflated Type-I error rate for this approach.

Simulations from the unbalanced sample size scenarios show the same conclusions with a magnified support for the Empirical Bayesian approach (Results not shown). Comparisons of the performances of the methods or a non-rare scenario was also provided in Table 3, where we conclude that the empirical Bayesian approach is the most conservative in retaining Type-1 error rate and has competitive performance in estimating the targeted Odds ratio.

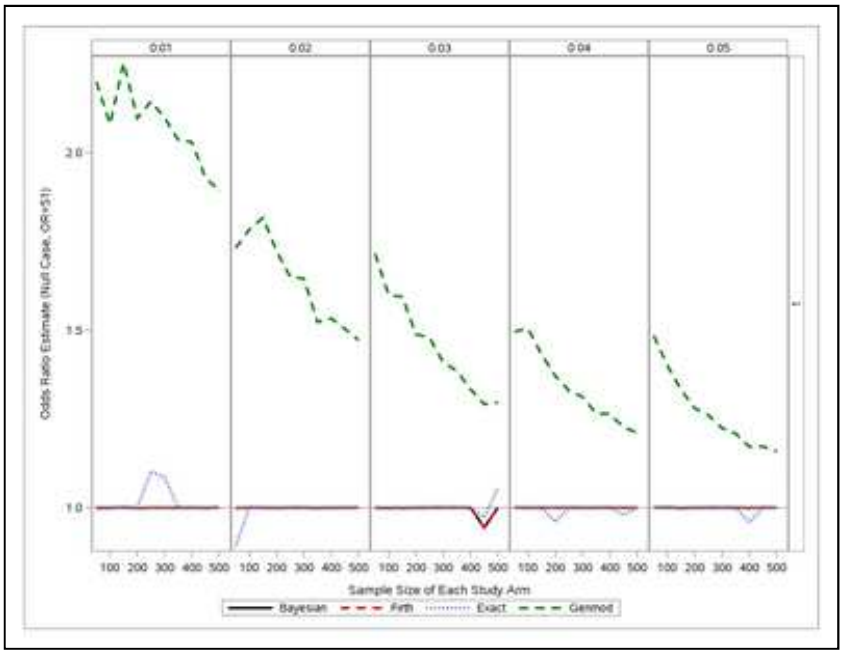

FIGURE 1: Odds Ratio estimate from the competing methods when the true Odds Ratio=1.

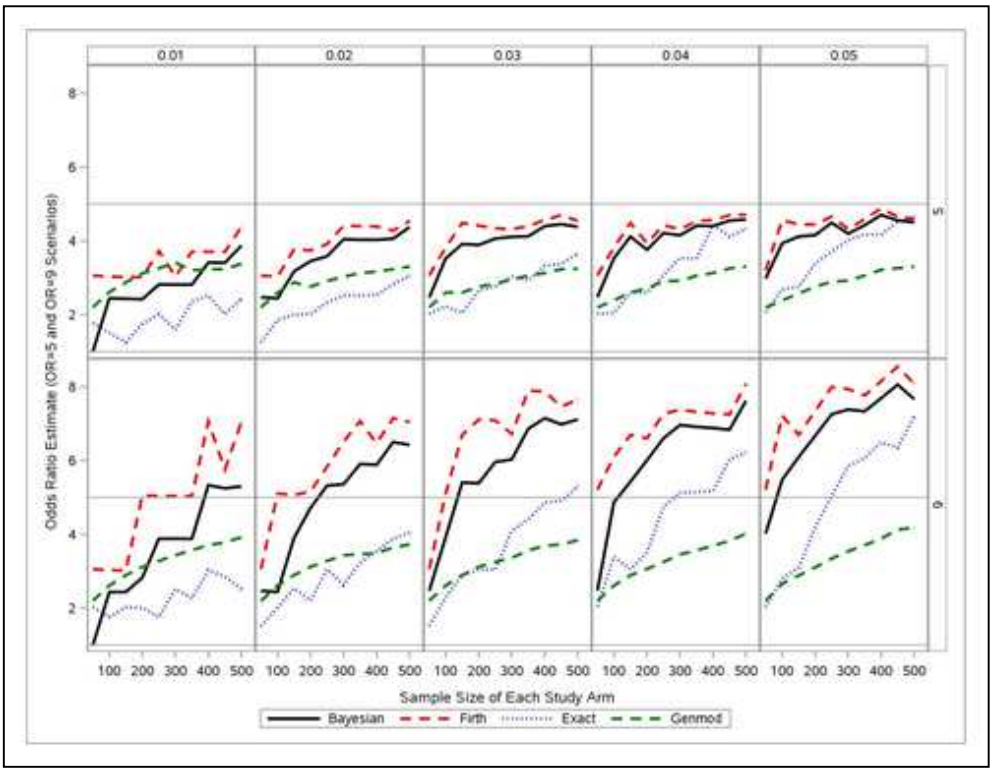

FIGURE 2: Odds Ratio estimate from the competing methods when the true Odds Ratio is 5 or 9.

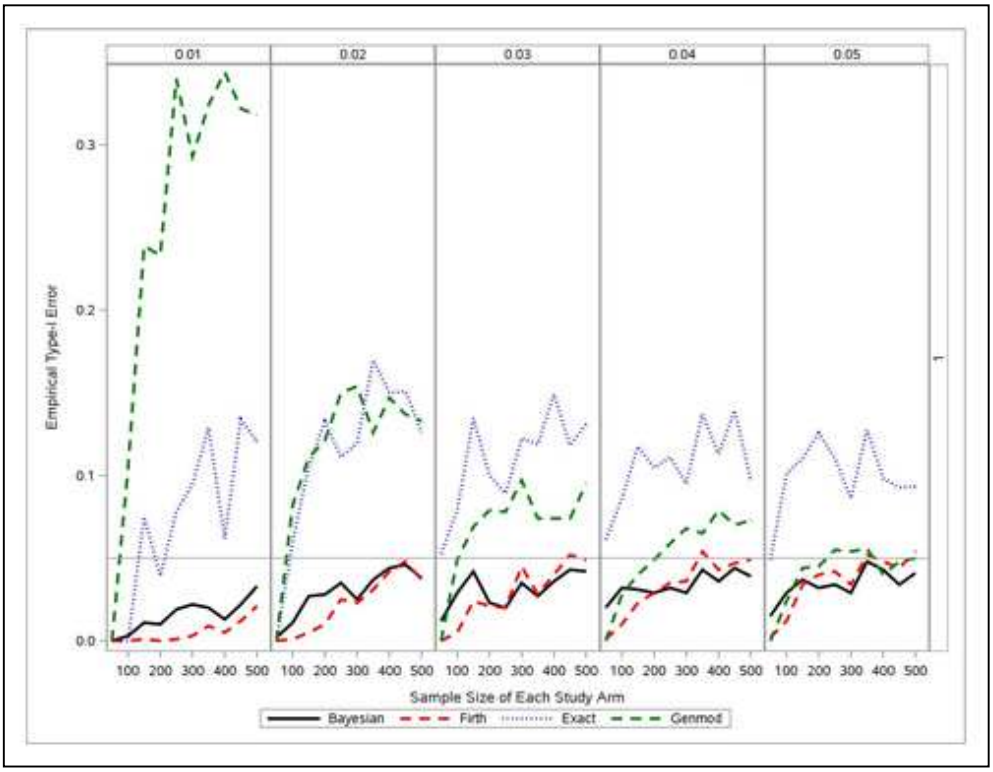

FIGURE 3: Type-1 Error Rate of the competing methods when the true Odds Ratio=1.

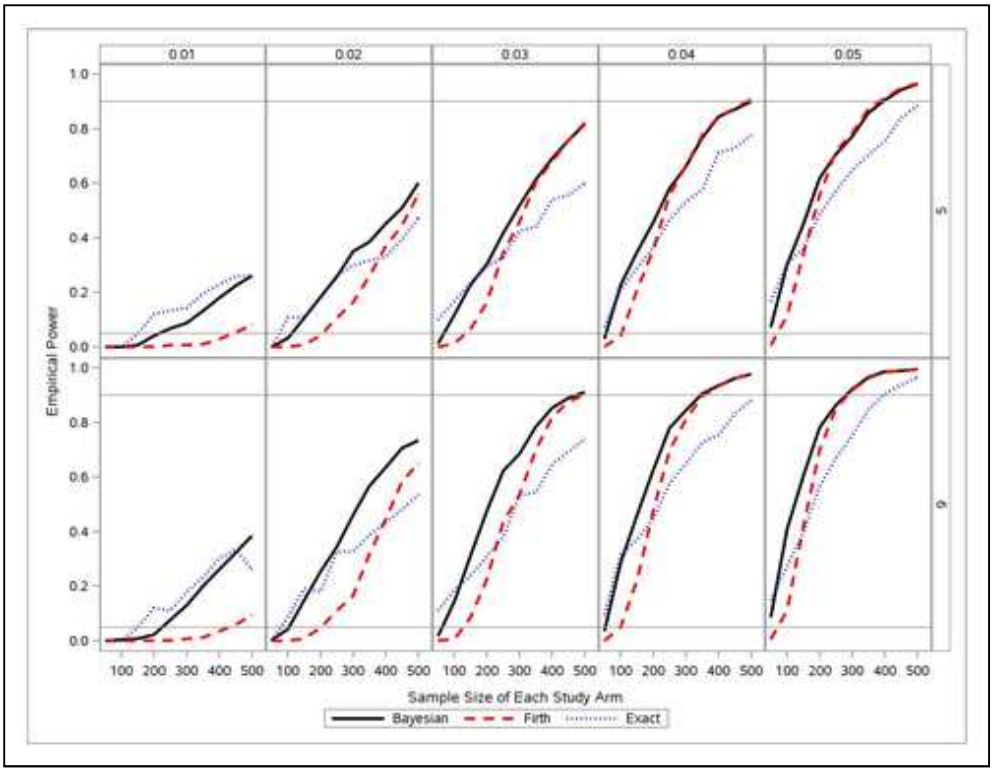

FIGURE 4: Empirical Power of the competing methods when the true Odds Ratio is 5 or 9.

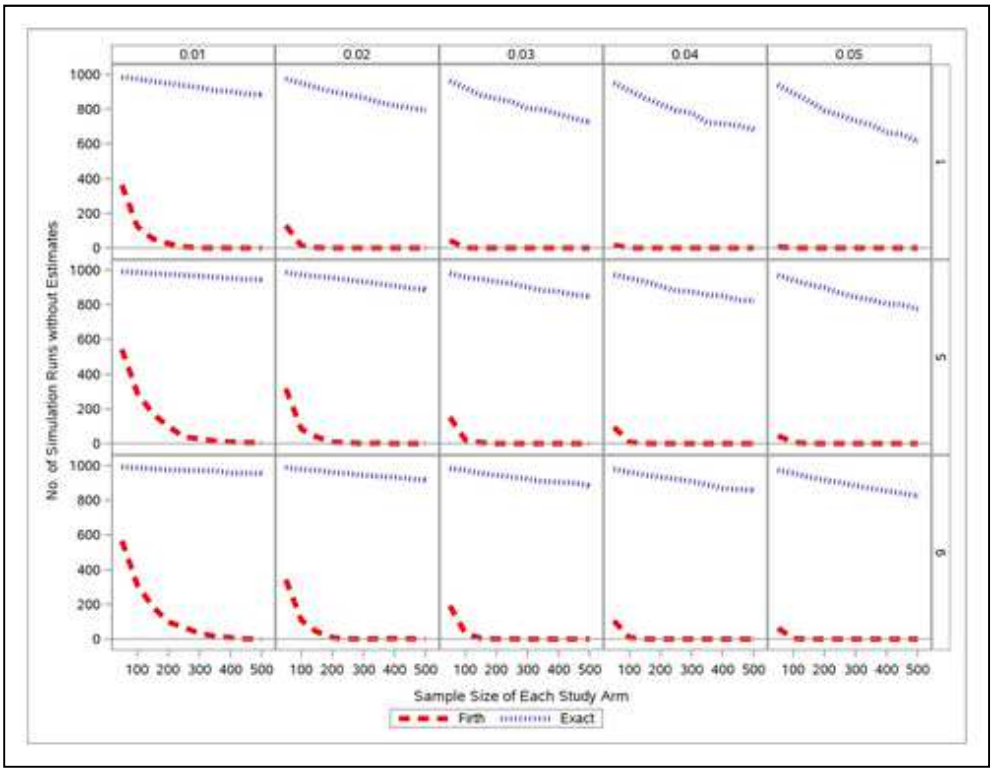

FIGURE 5: No. of simulations where estimation was not possible for Firth and Exact approaches.

TABLE 3: Power of the competing methods for a non-rare scenario with success probability of 0.90, sample size in each arm of 500.

	True OR	Or Estimate	Power	Method
	1	1.000	0.037	New Bayesian Method
	1	1.000	0.053	Firth Approach
	1	0.990	0.057	Exact Logistic Approach
	1	0.986	0.045	Bayesian Logistic through GENMOD
	5	4.910	1.000	New Bayesian Method
	5	4.921	1.000	Firth Approach
	5	4.970	1.000	Exact Logistic Approach
	5	4.128	1.000	Bayesian Logistic through GENMOD
9	9.229		1.000	New Bayesian Method
	9	8.952	1.000	Firth Approach
	9	9.131	1.000	Exact Logistic Approach
	9	6.517	1.000	Bayesian Logistic through GENMOD

APPLICATION

In CANDLE study, which is a birth-cohort study, 1020 mother-child dyads had 2-year cognitive assessment.[8] Of these, the alcohol consumption during pregnancy by race is given in Table 4. We wish to compute the odds ratio of alcohol consumption during pregnancy in African American (AA) cohort versus Caucasian (CA) cohort.

TABLE 4: Real-data application from the CANDLE study.

	Alcohol	No Alcohol	All Patients
African Americans	34	623	657
Caucasians	52	311	363
All Patients	86	934	1020

Table 5 presents the Odds Ratio estimates comparing African Americans with Caucasians for the likelihood of alcohol consumption.

TABLE 5: Odds Ratio estimates comparing African Americans with Caucasians for the likelihood of alcohol consumption.

Method	OR	95% CI	P-value
Direct Estimate	0.326	0.207-0.514	<0.0001
Firth	0.328	0.209-0.515	<0.0001
Exact Logistic	0.327	0.201-0.525	<0.0001
Bayesian Logistic through GENMOD	0.503	0.203-0.749	0.0007
Empirical Bayes Estimate	0.329	0.208-0.513	<0.0001

We see that our Empirical Bayesian approach has the narrowest confidence interval followed by Firth and Direct approaches. Bayesian Logistic approach through GENMOD has the widest interval and an unreasonable Odds Ratio estimate quite far from the other four, with the least significance due to bigger standard error.

DISCUSSION

In this research, we proposed a Bayesian approach to estimate the Odds Ratio in rare event and zero-event cases and through extensive simulations, we have shown that our Bayesian approach is able to estimate an Odds Ratio even when one or both arms of the study has zero events, more powerful and has narrower confidence interval compared to its commonly used counterparts.

Being a Bayesian approach, it is more computationally expensive as expected; however, in today's computational power, this limitation becomes burdensome only when high number of Odds Ratio estimations are desired. We compared all these competing approaches regarding the CPU time on a Windows-7 PC with Intel® Core™ i7-4790 CPU @ 3.60GHz with 16 GB RAM on our sample data. Our approach took 13 cpu seconds, while the Genmod approach took 30 cpu seconds for the same problems. The other

APPENDIX: We present our SAS Macro program, which can be implemented using the sample data we shared above.

```
%macro orbayes (n1=, n2=, n11=, n21=, postN=1000000, gsize=1000, outname=pvalue, rseed=123);
*** N1: Sample size from Population-1
*** N11: Number of successes (i.e., events) from Population-1
*** N2: Sample size from Population-2
*** N21: Number of successes (i.e., events) from Population-2
*** postN: Number of samples drawn from the posterior distribution
*** gsize: Size of the clusters
*** Outname: Name of the output data to be captured
*** Rseed: Random seed that will allow to replicate the same simulation if needed
***;
data postsamp; n1=&n1; n11=&n11; n2=&n2; n21=&n21; call streaminit(&rseed);
do id=1 to &postN;
p1=round(rand('beta',n11+1, n1-n11+1),0.00000000001);
p2=round(rand('beta',n21+1, n2-n21+1),0.00000000001);
logor=log((p1/(1-p1))/(p2/(1-p2)));
output; end; run;
data postsamp; set postsamp; ngroup=ceil(_n_/&postN*&gsize); run;

proc univariate data=postsamp noprint; by ngroup; var logor;
output out=samplebound pctlpre=P_ pctlpts= 2.5 50 97.5; run;
data samplebound; set samplebound; if p_2_5>0 or p_97_5<0 then significant=0;
else significant=1; run;

proc sql; create table &outname as
select distinct &n1 as n1, &n11 as n11, &n2 as n2, &n21 as n21, exp(mean(p_50)) as or_median, exp(mean(p_2_5)) as or_LB,
exp(mean(p_97_5)) as or_UB, mean(significant) as pvalue from samplebound; quit;
proc sql; drop table postsamp, samplebound; run;
%mend;
%orbayes(n1=30, n2=36, n11=6, n21=0, outname=pvalue, postN=10000000, gsize=10000, rseed=1236);
proc print data=pvalue; run;
```

two approaches namely, Firth and Exact Logistic, naturally took much shorter cpu time, 0.09 and 0.07, respectively.

There are a couple of limitations to our Bayesian Odds Ratio estimation approach. Currently, the Empirical Bayesian method is only available in SAS and we plan to generate an R-version of this procedure in near future. In its current algorithm, no stratification is built-in as a possibility and the method should be expended to have such flexibility to control for a factor of interest in estimating Odds ratios. In addition, currently, the Bayesian algorithm assumes Uniform priors for the probability of events and more flexible priors like Beta-distribution can be explored to increase the power when such priors can be justified from earlier studies. Like in any other modeling framework, just having Odds ratios would not be sufficient and any methodology should be extended to include other control variables and covariates to test whether or not the primary comparison, say, between two treatment arms, remains significant in the presence of other known factors. Therefore, we plan to expand our Bayesian method such that a type of Bayesian Logistic Regression modeling framework can be established.

Conflict of Interest

Authors declared no conflict of interest or financial support.

Authorship Contributions

The Author Mehmet Koçak took 100% active role in article writing and applying statistical methods and has the 100% responsibility.

REFERENCES

1. Albert A, Anderson JA. On the existence of maximum likelihood estimates in logistic regression models. Biometrika 1984;71(1):1-10.
2. Harrell FE Jr, Lee KL, Matchar DB, Reichert TA. Regression models for prognostic prediction: advantages, problems, and suggested solutions. Cancer Treat Rep 1985;69(10):1071-7.
3. So Y. A tutorial on logistic regression. Proceedings of the Eighteenth Annual SAS Users Group International 1993;1(1):1290-5.
4. Agresti A. An Introduction to Categorical Data Analysis. 1st ed. New York: John Wiley & Sons, Inc; 1996. p.312.
5. Lesaffre E, Albert A. Partial separation in logistic discrimination. J R Stat Soc Series B 1989;51(1):109-16.
6. Zorn C. A solution to separation in binary response models. Polit Anal 2005;13(2):157-70.
7. Firth D. Bias reduction of maximum likelihood estimates. Biometrika 1993;80(1):27-38.
8. Tylavsky FA, Kocak M, Murphy LE, Graff JC, Palmer FB, Völgyi E, et al. Gestational Vitamin 25(OH)D Status as a Risk Factor for Receptive Language Development: A 24-Month, Longitudinal, Observational Study. Nutrients 2015;7(12):9918-30.