JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4368-4379
https://publishoa.com
ISSN: 1309-3452

An Empirical Study on Imbalanced Samples and Presence of Zeros in Clinical Trials

S.Sumathi[1*], B.Senthil Kumar[2]

[1*]Research Scholar, PG and Research Department of Statistics, Periyar EVR college (Autonomous), Trichy, India.

[2]Assistant Professor, PG and ResearchDepartment of Statistics, Periyar EVR college(Autonomous), Trichy, India.

Corresponding Author email: s.sumathi0909@gmail.com

Received: 2022 March 15; **Revised:** 2022 April 20; **Accepted:** 2022 May 10

ABSTRACT

Meta-analysis is a statistical tool in medical research to analyse exposed effect by estimating the binary outcomes from multiple clinical studies. The main objective of this paper is to study the sensitivity of parameter in a Meta-analysis from imbalanced dataset of clinical trials. An empirical study on data from 24 clinical trials has been analysed using Bayesian random effects model. The parameter of interest is the risk difference between the exposed group and the unexposed of balanced and imbalanced groups. Presence of one or more zeros in a study could create a major impact in variability between studies. This study has focussed of extent of sensitivity in the estimation of point and confidence intervals.

Keywords: *Risk difference, Balanced and Imbalanced group, Bayesian REM, Meta-analysis, Point and Interval Estimate.*

1. INTRODUCTION

The systematic overview, otherwise known as a meta-analysis, is a statistical process in which the results of multiple independent studies are merged, in order to evaluate the extent of Variability, as it plays an important role in clinical studies.Let X_1 and X_2 denote two categorical response variables, X_1 with Y_1 categories and X_2 with Y_2 categories leading to Y_1Y_2possible combinations. A rectangular table having Y_1 rows for categories of X_1 and Y_2 columns for categories of X_2 displays this distribution. The joint distribution determines the association(magnitude) between two categorical variables. The cells of the table represent the Y_1Y_2 possible outcomes. The cells inlcude frequency counts of results for a sample, such table is called a contingency table or cross-classification table. Usually referred as 2x2 table. The outcome in every independent study is a binary variable and it can be viewed as a two-by-two contingency table, with each cell corresponding to counts of events in separate groups. For example, participants assigned to treatment(X_1) and control arms(X_1) of aninterventionstudy.In clinical studies, association between X_1 and X_2 is analysed i.e., proportion of success in row1 to that of proportion of success in row 2. The interest in this empirical study is estimating Risk Difference in the success rates of the 2 rows that is often used as the measure of effect in practice.More than one 2x2 tables are grouped and analysed.Merging this sort of data can be problematic when zero event occur either in one or both arms of a study.The sparsity in this study could create an impact in estimating the summary measures, computational complexity and asymptotic approximations, which assumesof having an observation is theoreticallynot possible and does not contribute in the mechanism of estimation or in fitting appropriate models. It’s quite common to remove such studies(Liao, 1999) as they do not provide reasonable information on the magnitude of the treatment effect (Sweeting et al., 2004) or another common remedial measure is, the studies require continuity correction(Skene and Wakefield, 1990), which may influence the results.Another important aspect has been noticed when merging 2x2 tables is, thesample size. It plays an important role in clinical studies. For example, if number of patients in control arm is 2 times of that of treatment, then this might either create an impact in finding the association between exposed and unexposed group or model may be affected.The main interest of this empirical study is to compare the point (RD) and interval estimate of a Balanced and an Imbalanced group of a sparse dataset, which is considered as the data characteristics, which has less attention in Bayesian studies. In Bayesian inference, priors isassignedto all unknown parameters p(θ), the likelihood is defined for the data given the parameters f(X|θ) and posterior distribution π(θ|X) of the parameters is determined given the data using Bayes' theorem,$\pi(\theta|x) = \frac{p(\theta)f(x|\theta)}{\int p(\theta)f(x|\theta)d\theta}$ assuming θ as a continuous random variable. This study describes about a fully Bayesian random-effects model and how other issues to be handled within aintegrated framework, and alsoregarding how graphical modelling techniques can contribute valuable perception.

The entire exercise has been carried out in extracting the datasets from clinical trials.Either one or more than one tables from these studies are chosen for analysis. All these single tables are combined and grouped based on characteristics of data such as balanced and imbalanced, also number of zeroes in the study. Twenty-four 2x2 tables has been collected from Eleven published clinical studiesand analysed, where each table is considered as a Dataset and handled individually.

Datasets have been named as D1, D2, …etc. Where D1 to D6 is considered as Balanced group and D7 to D24 is considered as imbalanced group.Each of these group includes the study that has one or more than one zero.Details of the studies are as follows: – Kishore study has fourteen single tables where only two tables has been taken and named as D1 and D8, Skene and Wakefield(1990) has a total of eight tables from which four tables has been taken and named as D2, D7, D17, D18, Agrestihas eight individual studies, three studies have been considered and named as D3, D13, D22,Efron 1996 has forty-one tables, four tables has been extracted and named asD4, D5,D6,D20,Carlin(1992) consists of twenty-two tables and only one table has been considered, it is D9.Hardyhas nine studies, one study has been considered as D10,Sweeting., et al(2004) has thirty studies, four has been drawn out and named as D11, D19, D21, D24,Smith., et al (1995) comprise of twenty – two studies, one has been drawn out and has been named as D12,Warn.,et al has forty-six tables, only one satisfies the objective of this study, so one study has been wrenched out and named as D14,Tian et., al(2007)has forty-eight tables, two studies has been taken and named as D15, D23, Cochran has four tables from which only one table has been considered for evaluation, it isD16.

Table 1 describes the dataset that has been wrenched out from published clinical studiesbased on the data characteristics and grouped accordingly, which fits the objective of this study. This is a single table analysis, where each table is a study. Here the table comprises of Nine columns, where Data no refers to the individual table that has been extracted from clinical studies, ai and bi refers to treatment arm towards exposure and non-exposure, ci and di refers to control arm towards exposure and non-exposure respectively. The ratio column is found by dividing the sum of ai, bi and sum of ci,di. If the ratio is equal to 1, the study is said to be balanced. If the ratio is less than 1 we then find the inverse of the current ratio to form the new ratio, as it is always majority class to the minority class. However, the group is still considered to be imbalanced group. This is done is Rn1n2 column. BorUB mentions about the characteristics of the data based on ratio, where B refers to Balanced and UB refers to Imbalanced group. The last column shows the number of Zeros (0 – No zero, 1 – One zero and 2-Two zeros) in each study.

Table 1Grouping of Datasets from multiple Clinical studies on Data characteristics

DataNo	**ai**	**Bi**	**ci**	**Di**	**Ratio**	**Rn1n2**	**BorUB**	**Zerotype**
D1	9	14	5	18	1	1	B	0
D2	14	5	7	12	1	1	B	0
D3	3	3	0	6	1	1	B	1
D4	3	9	0	12	1	1	B	1
D5	0	6	6	0	1	1	B	2
D6	0	34	34	0	1	1	B	2
D7	1	6	2	7	0.7778	1.2857	IB	0
D8	11	25	10	27	0.973	1.0278	IB	0
D9	28	223	12	110	2.0574	2.0574	IB	0
D10	21	364	17	117	2.8731	2.8731	IB	0
D11	0	150	15	6821	0.0219	45.5733	IB	1
D12	1	607	37	6142	0.0984	10.1628	IB	0
D13	6	11	0	1	17	17	IB	1
D14	0	8	2	13	0.5333	1.875	IB	1
D15	14	11	23	0	1.087	1.087	IB	1
D16	2	29	0	11	2.8182	2.8182	IB	1
D17	2	561	0	142	3.9648	3.9648	IB	1
D18	17	16	0	4	8.25	8.25	IB	1
D19	2	0	5	0	0.4	2.5	IB	2
D20	0	1128	0	137	8.2336	8.2336	IB	2
D21	0	39	0	78	0.5	2	IB	2
D22	1	10	0	1	11	11	IB	1

D23	0	9	0	16	0.5625	1.7778	IB	2
D24	0	676	0	225	3.0044	3.0044	IB	2

2. METHODS AND MEASURES

Association measure used in this study is Risk Difference, a descriptive measure for comparing groups on binary responses. In a 2x2 table, RD is defined as a difference in proportions and it lies between -1 and +1. If RD equals zero, then the 2 responses are statistically independent.The point estimate is the difference in sample proportions, as shown by the following equation:

RD = p_1-p_2

The proportions of sample are evaluated by considering the ratio of the number of successes to the sample size (n) in every group:

$$p1 = \frac{X1}{n1} \text{ and } p2 = \frac{X2}{n2} \qquad (1)$$

whereX_1 is success in row 1, X_2 is the success in row 2, n_1 and n_2 are row 1 total and row2 total respectively.Risk difference is also referred as attributable risk, and when expressed in terms of percentage it is also referred to as attributable proportion. Risk difference is used to valuate the risk in the exposed group that is attributable to the exposure. Risk difference can be directly illustrated even without knowing the risk of the control group as it focuses on the absolute effect.For example, in a simple Bayesian binomial model for a zeroevent trial with sample sizes n_1 and n_2, parameters p_1 and p_2 will yield a posterior mean estimate of risk differencep_2- $p_1$accompanied by uniform prior on event rates, with 95% quantile-based credible interval.Bayesian estimation of risk difference for rare events is focused throughout the study. Agresti (2003) describes more details and its advantages so as to appreciate it as a desired summary measure in REM with binary data.

In this paper, we explore the various types of priors that are used in Bayesian estimation and analyse the risk difference formulated on the information collected from these studies and also has focussed on the extent of sensitivity in the estimation of point and confidence intervals.Bayesian analysis requires computing belief of functions of random quantities as a basis for conclusion, where these quantities may have posterior distributions.Inferences throughout are based on Rusing Monte Carlo Markov Chain. R is a statistical software package, and hence the same analysis can be easily implemented in another problems.

3. RANDOM EFFECT MODEL

Meta-analyses uses REM, random effects model (Smith et al 1995) allows for differences in the treatment effect from study to study. It is a statistical procedure to synthesis the outcome of individual studies in order to estimate the study effect and assess whether study effects are similar enough to be combined. All reviews have been narrated, but the narrative review is mostly subjective since different experts can come to different conclusions, and becomes difficult when there are more than a few studies involved.It is necessary to understand the sources of variability between study when making conclusions about the population. Such models are of considerable scientific interest and closely resemble the statistical principles of meta- analysis.In Bayesian model, parameters are also random variables and the Beta-Binomial model is described here. Bayesian method requires appropriate constants for parameter values, this helps to derive the shape and location of the Beta distribution by comparing the estimates.

Let us consider, Binomial distribution as the likelihood for the success factor.Then

X_1~ Binomial (n_1,θ_1)

X_2~ Binomial (n_2,θ_2)

Assuming suitable priors for two parameters θ1 and θ2,

θ_1~BETA (α_1,β_1)

θ_2~BETA (α_2,β_2)

we get a posterior for the proportions (RD= θ_1-θ_2) is Beta Distribution with updated parameters.

4. DATA ANALYSIS

Table 2 deals with the problemson estimating the risk factor between of Twenty-four datasets that has been extracted from multiple clinical trials.A comparison study of point and interval estimate has been done between balanced group D1 to D6 and imbalanced group D7 to D24.In order to analyse the variation between balanced and imbalanced group, priors for Beta distribution has been fitted with symmetric parameters (0.001,0.001) and unsymmetric parameters (5,1).

Table 2Estimated Point and 95 % confidence interval limits for risk difference from the 2 x 2 contingency tables considered in the study. Priors for beta distribution is p_1 ~ Beta(0.001,0.001), p_2 ~ Beta(0.001,0.001) and p_1 ~ Beta (5,1), p_2 ~ Beta(5, 1)

	p1 ~ Beta (0.001,0.001) **p2 ~ Beta(0.001,0.001)**			**p1 ~ Beta (5, 1),** **p2 ~ Beta(5, 1)**		
DATANO	**ESTIMATE**	**LOWER**	**UPPER**	**ESTIMATE**	**LOWER**	**UPPER**

D1	0.1740	-0.0820	0.4260	0.1380	0.1120	0.3800
D2	0.3680	0.0670	0.6350	0.2800	0.0230	0.5240
D3	0.5000	0.1470	0.8530	0.2490	0.1330	0.5960
D4	0.2500	0.0600	0.5180	0.1660	0.1420	0.4590
D5	-0.9950	-1.0000	-0.9360	-0.5000	-0.7820	-0.1720
D6	-0.9990	-1.0000	-0.9890	-0.8500	-0.9430	-0.7220
D7	0.0350	-0.1680	0.2390	0.0320	0.1690	0.2300
D8	-0.0800	-0.4410	0.2950	-0.0060	-0.3610	0.3530
D9	0.0130	-0.0560	0.0770	-0.0040	-0.0780	0.0650
D10	-0.0720	-0.1370	-0.0160	-0.0910	-0.1590	-0.0300
D11	-0.0040	-0.0070	0.0000	0.0030	0.0040	0.0120
D12	-0.4390	-0.6320	-0.2530	-0.3520	-0.5370	-0.1750
D13	-0.1330	-0.3400	-0.0170	0.0240	-0.2830	0.3460
D14	0.0640	0.0080	0.1730	-0.1060	-0.3600	0.1260
D15	0.0040	0.0000	0.0100	-0.0210	-0.0570	0.0040
D16	0.5150	0.3470	0.6810	0.0640	-0.2640	0.3950
D17	0.0890	0.0020	0.3060	-0.3600	-0.7040	0.0580
D18	0.3510	0.1450	0.5870	-0.2360	-0.5680	0.1680
D19	-0.0020	-0.0030	-0.0010	0.0290	0.0080	0.0620
D20	0.0000	0.0000	0.0000	0.1070	0.1740	0.3960
D21	0.0000	0.0000	0.0000	0.0510	0.0430	0.1640
D22	-0.0080	-0.1750	0.0590	-0.0340	-0.3450	0.2330
D23	0.0000	0.0000	0.0000	-0.0140	-0.0370	0.0020
D24	0.0000	0.0000	0.0000	-0.0310	-0.0670	-0.0070

4.1. Summary

Data sets that are extracted from various published literature is analysed in order to understand the data characteristics and the parameter sensitivity of a balanced and an imbalanced group.Its been observed that practical data sets are more essential and relevant for the current study, which helps in making out the practical objectives and the way they mutate to statistical objectives and their influence on subsequent analysis. Hence, it is very important to consolidate all data sets collected from the extensive literature. This helps to illustrate the practical concerns and to translate into the statistical analysis. Process of data collection exercise is based on twocomponent, one is the sample size of the study and secondly the number of zeros present in the study.All these studies havebeen grouped accordingly and analysed. A beta distribution has been fixed as prior with parameterα and β, where α and β have been assigned values to find the behaviour of the study. Twenty-five different combinations of parameters(α and β) values has been analysed, which includes the symmetric[(0.1,0.1),(0.5,0.5),(1,1),(5,5)…)] and non-symmetric parameter values[(5,1),(1,0.1),(0.001,1)…].This work has been carried out in order to analyse the sensitivity of parameter, Point estimate and Interval estimate between the two groups.

A forest plot is an essential tool to sum up information on individual studies addressing the same question.It gives a visual suggestion about the study and show the estimated effect in one figure. A forest plot arrays point estimates and interval estimate (e.g., 95% CI) represented by whiskers for multiple studies in a horizontal orientation. A vertical line is typically plotted at the null hypothesis, with the statistical importance of an individual point and whiskers compared to that reference line. It summarizes Impact of the RD estimate between study for the Imbalanced data with assorted parameter values.

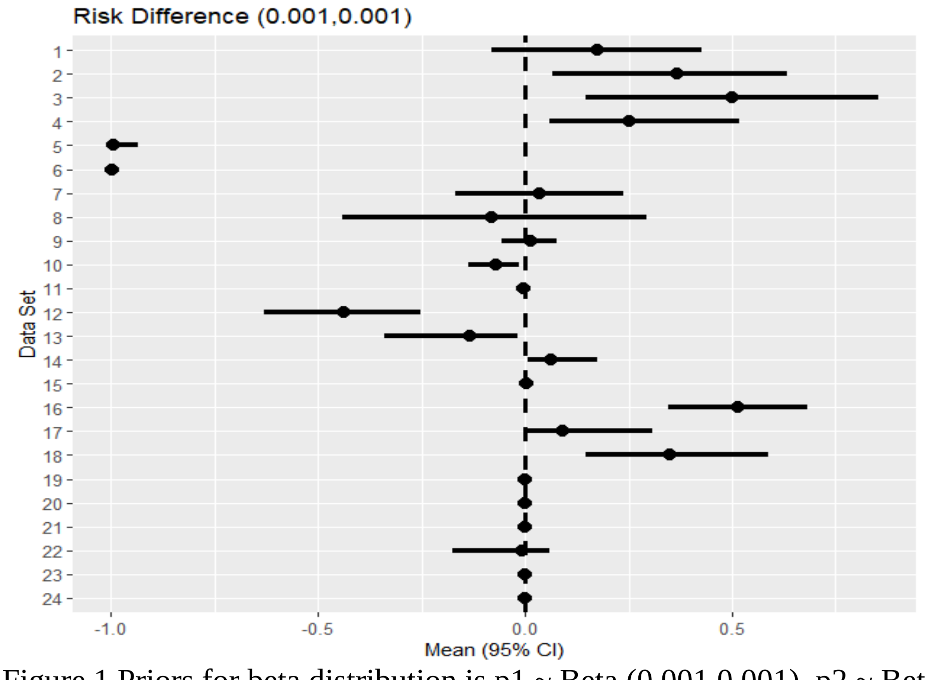

Figure 1 Priors for beta distribution is p1 ~ Beta (0.001,0.001), p2 ~ Beta(0.001,0.001)

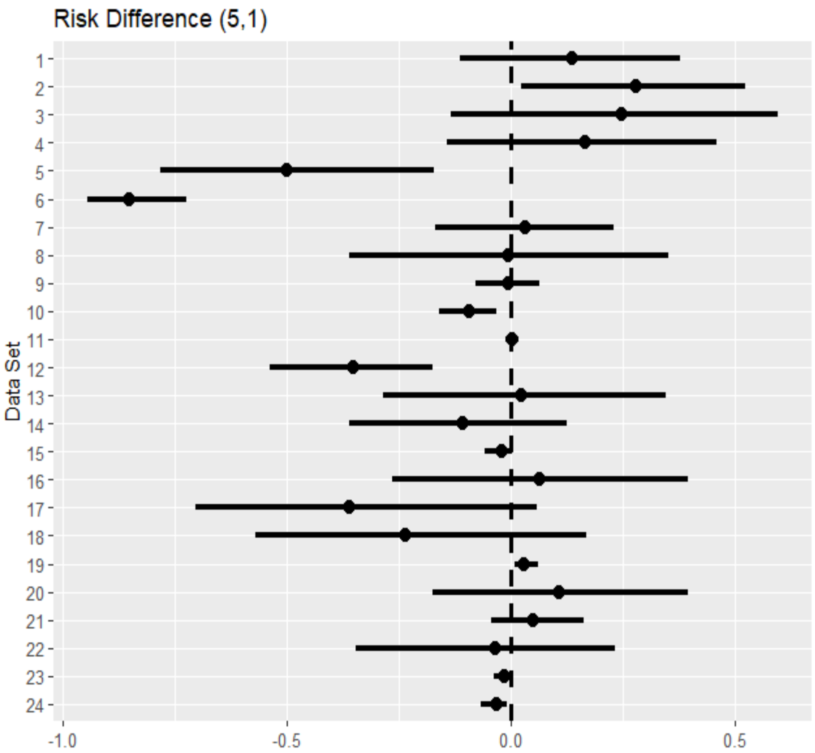

Figure 2 Priors for beta distribution is p1 ~ Beta (5,1), p2 ~ Beta(5,1)

4.2. Balanced group

InFigure1 and 2, D5 and D6noted to be statistically dependent towards the unexposed for parameter values (p1-0.001, p2-0.001) and (p1-5, p2- 1).Interval estimates of D5 and D6 is observed to be augmented for non-symmetric parameter, this may be due to presence of more than two zeros in the study. D1 may be statistically independent for both symmetric and non-symmetric parameter values, as the interval estimates lies between (-1,+1). D3 and D4 is observed to be biased towards the exposed group for symmetric parameter value, where as it is observed to statistically independent for non-symmetric parameter. D2 shows not much differenceas there may not any appreciable difference in the point estimates.

4.3. Imbalanced group

D10 and D12 are perceived to be biased towards unexposed group, shows not much difference in point and interval estimates. D7 and D8 has a positive and negative point estimate and may be statistically independent as their confidence interval crosses the line of null effect. Though there is no zeros present in D9, there is a small variation in point estimate due to parameter values. It shows a positive RD for greater parameter (5,1) and a negative RD for smaller parameter values (0.001,0.001). In Both fig 1 and fig 2, D11 is observed to be on line of null effect, but as far as point estimate is concerned it yields a positive RD for non-symmetricparameters with interval estimates (0.040, 0.120) and negative RD for symmetric parameters with interval estimates (-0.007, 0) due to zero in one arm of the study.Therefore we can conclude that D11 may be statistically independent. D13 and D14 exhibits alternately greater than and less than 1 for symmetric and non-symmetric parameters and shows not much difference in point and interval estimates, so it is likely to independent. Like D11, D15 may be statistically independent as the interval estimate lies between (0,0.01) and biased towards the unexposed group for parameter values (0.001,0.001). Also the interval estimate of D15 expands for non-symmetric parameter (5,1) due to zero in one of the cell in the respective study. D16, D17 and D18 are biased towards the unexposed group for (0.001,0.001) and crosses the line of null effect for parameter values (5,1). Here we can observe that the number of zeros present in one arm of the study creates an impact in the result. so we can conclude that the studies D16, D17 and D18 may be statistically independent.Though D19 has zeros in two arms of the study, there is no much notable difference in the point and interval estimates. D22 intimates that the study is biased towards unexposed group and interval estimates widens for non-symmetric parameters. D20, D21, D23 and D24 are statistically independent with point estimate 0 and interval estimates (0,0) for symmetric parameters(0.001,0.001). The same shows a sudden transformation in point and interval estimates (D20: 0.1070, (0.1740, 0.3960)), (D21: 0.0510, (0.0430, 0.1640,)), (D23: -0.0140, (-0.0370, 0.0020)) and (D24: -0.0310, (-0.0670, -0.0070) for non-symmetric parameter values (5,1). All these studies (D20, D21, D23 and D24) can result to be statistically independent, this is mainly due to the influence of more than one zeros in the study.Twenty-five parameter (p1,p2) combinations have been analyzed in this study, only two parameter combinations ((0.001,0.001) and (5,1))have been displayed in this paperdue to dearth of place.

5. CONCLUSION

The main objective of this paper isto analyse association measure of a 2 x 2 categorical data using Bayesian inference, that has received active research attention but mainly in classical procedures. This studysummarizes the extent of sensitivity in estimation point and confidence interval of association measure such as risk difference, is usually difficult to establish, especially for sparse data or data with presence of sampling zeros. Attention has been paid in interpretation of summary measures.The study indicatesaboutthe difference in RD estimate and Confidence Interval, in order to provide a structure for the analysis of parameter sensitivity with distinct data characteristics such as sparseness in terms of zero and also group size.Behaviour of the studies is noted with assorted parameter values with appropriate priors using Bayesian inference.Studies with more than one zero could create a major impact in the results.This work has identifies plausible way of developing a comprehensive Bayesian procedures and its implementation for association measures. Further, this study has identified areas of application in medical / epidemiology / clinical trial to study various models for association measures and to make use of advantage of Bayesian approaches inbinary data.This empirical investigation provides a scope of dealing with the analysis of sparse (small or large)datasets through Bayesian framework.

STATEMENTS AND DECLARATIONS

Conflict of Interest

The authors declare that they have no conflict of interest.

Funding

No fund received for this project.

References

1. Agresti, A.: Categorical Data Analysis. John Wiley & Sons.(2003).
2. Berkey, C. S., Hoaglin, D. C., Antczak-Bouckoms, A., Mosteller, F., & Colditz, G. A.:Meta-analysis of multiple outcomes by regression with random effects. Statistics in medicine. 17(22), 2537-2550(1998).
3. Brown, L., & Li, X.: Confidence intervals for two sample binomial distribution. Journal of Statistical Planning and Inference. 130(1-2), 359-375(2005).
4. Carlin, J.B.: Meta-analysis for 2× 2 tables: a Bayesian approach. Statistics in medicine. 11(2), 141-158 (1992).
5. Efron, B.: Empirical Bayes methods for combining likelihoods. Journal of the American Statistical Association. 91(434), 538-550 (1996).
6. Gelman, A., Hwang, J., Vehtari, A.: Understanding predictive information criteria for Bayesian models. Statistics and computing. 24(6), 997-1016 (2014).
7. Gelman, A., Hill, J.: Data analysis using regression and multilevel/hierarchical models. Cambridge university press.(2006).
8. J. Sweeting, Michael, Alexander J. Sutton, and Paul C. Lambert.: What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. Statistics in medicine. 23(9), 1351-1375 (2004).

9. Lehmann, E. L., Casella, G.: Theory of point estimation. Springer Science & Business Media. (2006).
10. Robert, C.: The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer Science & Business Media.(2007).
11. Skene, A. M., Wakefield, J. C.: Hierarchical models for multicentre binary response studies. Statistics in Medicine, 9(8), 919-929 (1990).
12. Smith, T.C., Spiegelhalter, D.J., Thomas, A.: Bayesian approaches to random-effects meta-analysis: a comparative study. Statistics in medicine. 14(24), 2685-2699 (1995).
13. Subbiah, M.,Srinivasan, M. R.:Classification of 2× 2 sparse data sets with zero cells. Statistics & probability letters. 78(18), 3212-3215(2008).
14. Tian, L., Cai, T., Piankov, N., Cremieux, P.Y., Wei, L.J.: Effectively Combining Independent 2 x 2 Tables for Valid Inferences in Meta Analysis with all Available Data but no Artificial Continuity Corrections for Studies with Zero Events and its Application to the Analysis of Rosiglitazone's Cardiovascular Disease Related Event Data.(2007).
15. Tipping, M. E. (2003, February). Bayesian inference: An introduction to principles and practice in machine learning. In Summer School on Machine Learning (pp. 41-62). Springer, Berlin, Heidelberg
16. Warn, D.E., Thompson, S.G., Spiegelhalter, D.J.: Bayesian random effects metaanalysis of trials with binary outcomes: methods for the absolute risk difference and relative risk scales. Statistics in medicine. 21(11), 1601-1623 (2002).