JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4338-4355
https://publishoa.com
ISSN: 1309-3452

A Comparative Analysis to estimate Odds Ratio for K 2 x 2 Sparse Data Sets -Bayesian Modelling

S. Sumathi and B. Senthil Kumar*
PG Research &Department of Statistics,
Thanthai Periyar Government Arts and Science college(Autonomous), Trichy – 23.
(Affiliated to Bharathidasan University, Trichy - 24)
Corresponding Author: sumathi.phd19@gmail.com[1]
, senthilkumarbalan@gmail.com*

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

Abstract

An issue that exists during the meta-analysis process is that one or maybe more studies may have sparse data, such as no events in the treatment and control groups.Estimating Odds Ratio, one of the three association measures, will be challenging for such a dataset.In certain cases, it is standard procedure to either add a constant or remove the study to estimate the unknown values.This technique, however, is relied on the asymptotic characteristics of the estimates and may underperform for sample sizes. Another strategy would be to employ Bayesian approaches to have a better understanding of the problem. The main motive of this work is to perform a comparative study in analysing the robustness of highly imbalanced datasets with zero events for various values of K studies ($0 < k \leq 50$) in the dataset between the Binomial-Normal and Normal-Normal models.A Bayesian method with an appropriate prior might be a feasible option for dealing with sparse data, according to the findings of a comparative research.Both methodologies performed better in all sparse datasets and are suggested to be used in meta-analyses of similar scope.

Keywords:*Bayesian inference, Binomial-Normal, Continuity correction, Meta-Analysis, Normal-Normal, Odds ratio, Robustness, Sparseness.*

1. Introduction

Clinical trials are defined as systematic research studies conducted in humans to determine the safety and efficacy of emerging medicines, pharmaceutical combinations, medical treatments, technologies, or tactics. It encompasses human illness development, clinical mediation, the evolution of new technologies, epidemiologic and behavioural studies, as well as outcomes and health-care analysis. Clinicians use clinical trials to explore better methods to improve treatments and the quality of care for patients.

Clinical trials are used by professionals to properly assess if alternative medications are safe and effective, and if they outperform conventional treatments.The efficiency, execution, and interpretation of clinical studies are all influenced by the study design.Each study design has its own advantages and limitations. There may be a broad hierarchy of study designs that cannot

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4338-4355
https://publishoa.com
ISSN: 1309-3452

be implemented effectively to all study design styles.

The researcher defines the three elements of studies as follows: 1) description and measurement of exposure in two or more groups, 2) estimation of health outcome(s) in these same groups, and 3) statistical analysis performed between groups to determine potential associations between exposure and outcome.

The outcomes of clinical trials are typically presented in the form of a 2X2 table, also known as a contingency table (Green et al 2004). Contingency tables are a sort of classification table that has been used in statistics to describe associations between two categorical variables, variable X and variable Y. Contingency tables, also known as two-way tables, are a genre of table that is being used to summarise probability distributions including joint, marginal, and conditional distributions.

In clinical research, contingency tables are highly beneficial for evaluating observed effects with expected effects from a statistical model, and assessing adverse events between treatment arms.Row one of this 2 x 2 table represents the treatment group, while row two represents the control group (Agresti 2013). The results of a single study cannot be used to consider when deciding the efficacy of an intervention. As a consequence, multicentre statistical results are often collected across clinical domains and various groups using a qualitative approach. Here, each 2 x 2 table is considered a study, and each dataset is composed of k studies ($k > 0$).

Integrating the results of separate studies conducted at various centres but pertaining to the same aspect of research is a popular tactic in clinical trials.When the sample size is minimal in a two-by-two contingency table, there may be a number of cells with no observations, which is termed as sparse data. Adding a small constant to each cell of the observed table to obtain estimates of the unknown parameters is a common advice in traditional approaches in similar circumstances. However, because this method is based on the asymptotic properties of the estimates, it may underperform with small samples.

Bayesian techniques, on the other hand, provide an effective strategy for analysing the consistency of treatment effects across centres in a clinical trial dataset. It provides more flexibility in data analysis by analysing the underlying assumptions in the traditional approach as well as the small sample distributions of complex functions of parameters. The most important component of Bayesian technique, however, is determining an acceptable prior, which is necessary to arrive at a mathematically reasonable posterior.

Individual and overall odds ratios for sparse datasets have been shown using a hierarchical Bayesian (Chu et al 2018) model (Subbiah et al 2008). The main objective is to investigate how well severely imbalanced (Khalilia et all 2011) sparse studies perform in Binomial-Normal and Normal-Normal models for datasets with varying k values. A total of 83 clinical datasets were collected and analysed to validate the model. To enhance the interpretation of sparse data, the Bayesian method was developed.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4338-4355
https://publishoa.com
ISSN: 1309-3452

The most prevalent way for obtaining marginal densities from a Bayesian perspective is to employ simulation techniques, particularly the MCMC method. The introduction of high-speed computers and software has eased the above process and enable realisations from the joint posterior distribution of parameters of interest to be obtained (Higgins and Spiegelhalter, 2002).

The models and analytic approach for employing random effects models are described in Section 2. Section 3 focuses into the features of dataset definition and computation in R utilising Monte Carlo methods. Sec. 4 provides a discussion on the use of Bayesian techniques to sparse data, with a sidenote.

2. Models and Methods

The standard approach in a clinical study is to compare the treatment and control groups by monitoring the incidence within each group with regard to a certain quality of interest, such as the existence of a disease or not, as represented in a 2 x 2 table of the following form for the studies i.

Table 1 : Data Format

	Response		
	Yes	**No**	
Treatment	a_i	b_i	n_i
Control	c_i	d_i	m_i
	t_i	s_i	N_i

As per the above table, for study i,$n_i$is the number of cases in the treatment group and m_i is the number of cases in the control group, and ai and ci are the frequencies in the treatment and control groups, respectively, with regard to a particular attribute. $p_i^c = \frac{c_i}{m_i}$is the likelihood of an event occurring in the control group, whereas$p_i^T = \frac{a_i}{n_i}$denotes the probability of an event occurring in the treatment group. Then odds ratio is defined as,

$$r_i = \frac{p_i^T(1 - P_i^c)}{p_i^c(1 - p_i^T)}$$

The range of the odds ratio$r_i$is 0 to infinity. When the odds ratio is exactly 1, it implies that treatment exposure has no effect on the odds of control. An odds ratio greater than one indicates that there is a higher likelihood of treatment occurring as compared to exposure to control. An odds ratio of less than one indicates that exposure to treatment has a greater tendency to lead to control.

Prior distributions (Wu et al 2007) play an essential role in Bayesian modelling, particularly when the observational issue involves parameter space constraints. In Bayesian techniques, however, prior selection is critical. The Bayesian approach to statistical inference has acquired considerable interest in recent years in both practical and theoretical statistics. The objectives of this article is to make use of the inherent advantages of the Bayesian approach in statistical inference on ordinal characteristics with binary outcomes.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4338-4355
https://publishoa.com
ISSN: 1309-3452

Bayesian software packages primarily involve simulation techniques such as Markov chain Monte Carlo (MCMC)(Sorensen et al 2002) to generate a sample consisting of many draws from the intended posterior distribution.Individual, overall odds ratios, and heterogeneity for all eighty-three datasets were analysed in R (R Core Team 2016) using the Binomial-Normal (Pezeshk et al 2002) and Normal-Normal models (Frison et al 1992) to assess the robustness of unbalanced studies for a range of K studies.

2.1 Random effect model

The usage of the random effect model or REM (Riley et al 2011) in clinical data is still a matter of great debate in both standard (Agresti and Hartzel, 2000; Fleiss, 1986; Senn, 1998) and Bayesian approach (Abrams and Sanso, 1998; Gould, 2005; Skene and Wakefield, 1990). Maximum likelihood, constrained maximum likelihood, and Bayesian methods such as fully Bayesian approaches are all forms of random-effect models. The real impact size is expected to vary across studies under the random-effects model. Although these trials are not identical, they have enough in common to be included in the meta-analysis and have their data synthesised.

The random-effects model goes beyond chance to account for differences in effect magnitude across studies. This model suggests that there is no general effect size, but rather true study-specific effect sizes that are normally distributed around a mean (known as mean effect size) with variance that reflects the diversity of these effect sizes. Our objective was to evaluate the mean and variance of this distribution. The mean effect size is used to obtain the summary effect size. In the random effects model, each study offers a varied effect size due to both within- and between-study variances (the latter implying the degree of dissimilarity of the included studies), whereas in the fixed-effect model, only within-study variation is considered.

These models are quite scientifically interesting, and they closely resemble meta-analysis statistical principles. Several studies and/or meta-analyses have been published that address the theoretical, analytical, computational, and observational components of REM. REM is used in various fields for prospective research, despite the fact that medical, epidemiological, and wellness studies dominate this field.

In the case of a random-effects model, two methods are contemplated. The first is the fully Bayesian model (Binomial-Normal), which is based on the binomial method proposed by Smith et al. (1995), and the second is the Summary Statistics (Normal-Normal) model.

2.1.1. Binomial-Normal :-Fully Bayesian:

Consider the Binomial distribution as the likelihood of the success component. For the two rows of each table,

$$a_1 \sim \text{Binomial}\ (n_1, p^T)$$

$$c_1 \sim \text{Binomial}\ (n_2, p^C)$$

The number of people in the treatment group in row 1 is represented by a_1, which is calculated using n_1 events with a probability of p^T each. The number of people in row 2's control group is represented by c_1, which is calculated from n_2 events, each with a probability of p^C.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4338-4355
https://publishoa.com
ISSN: 1309-3452

Then we define $\delta=\text{logit}(p^T) - \text{logit}(p^C) = \log\left(\frac{p_i^T(1-p_i^C)}{p_i^C(1-p_i^T)}\right)$. This is the log-odds ratio, or the quantity of interest in log scale, for each of the k tables.For k tables, we additionally establish $\mu=\frac{\text{logit}(pT)+\text{logit}(pC)}{2}$.

This parameterization is the same as pT = $\text{logit}^{-1}(\mu+ \delta/2) = \frac{\exp(\mu+ \delta/2)}{1+\exp(\mu+ \delta/2)}$ and

$p^C = \text{logit}^{-1}(\mu- \delta/2) = \frac{\exp(\mu- \delta/2)}{1+\exp(\mu- \delta/2)}$

As a result, the model's second stage involves generating priors for μ and δ.

$\mu \sim N(\mu_i, \sigma_i^2)$ (i = 1,2,3,…,k)

$\delta \sim N(d, \tau^2)$

$\tau^2=\frac{1}{\eta} \sim$ Gamma (τ_1, τ_2)

The rate and shape parameters of the Gamma distribution are τ_1 and τ_2 respectively.Additionally, suitable priors for the scalar parameters μ_0, σ_0^2, d and τ_2 can be specified.Within variance isσ_0^2, and between variance is τ^2, which is a measure of variability among the k studies. Appropriate distributions are assumed for the next level of parameters $-\infty<\mu<\infty$ and $\tau^2>0$, which constitute the major focus of the study.

2.1.2. Normal – Normal:-Summary Statistics

Let's say there are k independent studies, each with an impact parameter Yi so that

$$Yi\sim N(\theta i , \sigma_i^2).$$

In addition, θi is supposed to come from a population with an impact size and is characterised as

$$\theta i\sim N(\mu,\tau^2).$$

σ_i^2 amounts to sampling variability in the i^{th} effect size estimate and τ^2 is the variability between effect size. In this model, σ_i^2 is assumed to be known in most of the occasion which is estimated from sample data.The sampling variability in the i^{th} impact size estimate is denoted by τ^2, while the variability between effect sizes is denoted by σ_i^2. σ_i^2 is considered to be known in most cases in this model, and it is calculated from sample data.

For Eighty-three clinical dataset

- k: Number of Studies in each dataset
- Yi: effect size – mean of k studies
- σ_i^2: Estimated sampling variability from k studies, using a simple formula
- θi, μ and τ^2 are to be estimated
- μ evaluates the overall mean and τ^2 measures the amount of heterogeneity.
- A value of τ^2 that is close to 0 implies low heterogeneity.

As a result, we model the sample mean as normally distributed with mean and variance as parameters (k studies). Then, in order to investigate the variability of the means, we form an effective level mean that follows a normal distribution. The required amount of heterogeneity estimator in the second level model is variance.

3. Dataset

3.1. Description

The complete dataset was culled from existing clinical research published in peer-reviewed journals between 1990 and 2021. The obtained datasets are organised in a Contingency table (Green et al 2004), a two-

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4338-4355
https://publishoa.com
ISSN: 1309-3452

by-two table that categorises two dichotomous variables, X and Y. In a study or experiment, X represents the treatment and control groups, while Y denotes the study's yes or no response. Each study in the dataset is represented by a table, and each dataset has k studies (k>0). Eighty-three datasets were retrieved, each having k 2x2 tables. These datasets were compiled in such a way that the study's cells had no zeros, one zero, two zeros, or more than two zeros.

Sample size is important when it comes to measuring significant measures. As a result, retrieved datasets have been classified as balanced (B) and unbalanced (IB) based on sample size. The major focus of this study is on evaluating the robustness of highly imbalanced sparse datasets for k studies. According to the number of studies (k) in each dataset, the extracted datasets have been divided into five groups: I, II, III, IV, and V.

The datasets have been designated D1, D2. . ., D83 after being collected from clinical studies, where studies in D1 are obtained from Agresti's book, D2 from Agresti 1992, and D83 from S.D. Walter1997.It is further categorized based on the number of studies, and as shown in Table 2, eighty-three datasets were sorted into five groups. The first column includes the names of the dataset's group. The second column displays the various k value ranges.

The dataset's columns are divided into four categories: B-0, B-1, IB ≥ 0, and IB≥ 1. B-0 refers to the number of datasets containing only balanced studies and no zeros. The number of datasets containing solely balanced studies with one or more than one zero is included in B 1. The number of datasets containing an imbalanced study with no zero is included in IB-0. IB ≥ 1 is the number of datasets containing an imbalanced study that contains one or more zeros.A few cells in Table 3 appear to be NA, indicating that no studies fall into that category. For instance, in group I,for k ≥ 0, B -0 indicates NA. Similarly, NA is present in a few cells in Table 2.

Table 2: Classification of the extracted datasets on characteristics

Group	k -Values	Categorisation of Datasets			
		B-0	**B ≥1**	**IB – 0**	**IB ≥ 1**
I	K<10	NA	D1, D40	D20,D3, D21, D23, D24, D25, D61, D30, D32, D33, D62, D8, D35, D67, D68, D69, D36, D11, D72, D74, D75, D76, D77, D78, D81	D58, D2, D59, D5, D31, D7, D63, D9, D10, D44, D45, D73, D12, D15
II	10≤K<20	D65	NA	D19, D28, D66, D37, D41, D50, D53, D54, D56	D60, D38, D64, D43, D47, D46, D48, D79, D51, D52, D83
III	20≤K<30	NA	D29	D4, D22, D71	D34, D26, D39, D70, D49, D80, D13, D14, D57, D55

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4338-4355
https://publishoa.com
ISSN: 1309-3452

IV	30≤K<40	NA	D27	NA	D42, D82
V	40≤K<50	NA	NA	NA	D18, D6, D16, D17

3.2. Analysis

As previously stated, the purpose of this study is to determine the reliability of the studies included in each dataset using a comparison study between groups. With the exception of the fact that all 83 datasets have been studied and analysed, it's hard to compare since the balanced group (B-0 and B-1) includes NA in the majority of the cells. As a result, our primary focus has moved to the highly imbalanced category with more than one zero (IB ≥ 1), despite the fact that datasets exist in all of the categories. As a result, the behaviour of severely imbalanced datasets is investigated over a wide range of k values.

This seems to be accomplished by either running a prolonged single MCMC chain or a huge number of shorter chains. Kernel density charts and other graphical tools can aid in analysing the convergence of MCMC chains. The primary challenge in MCMC is to identify the number of preliminary iterations M that should be discarded, and then to store every Kth result for the next N iterations.In the context of a single lengthy run of MCMC iterations, the methods for finding these numbers are explored (Rafter et al. 1992; Casella et al 1992).

The two models presented here are fully Bayesian, popularly known as the Binomial-Normal model (M1), and summary statistics, as well known as the Normal-Normal model (M2). These two models may be used to investigate the behaviour of balanced and unbalanced groups.Out of the eighty-three datasets, 41 fall into the IB≥ 1 group. There seem to be 14 datasets in Group I, 11 datasets in Group II, 10 datasets in Group III, 2 datasets in Group IV, and 4 datasets in Group V. Despite the fact that all forty-one datasets have been investigated and evaluated, just a handful have been shown here owing to space constraints.

3.2.1 Group I :K ≤ 10

Seven datasets exhibit variance between the two models M1 and M2 in group I, out of 14 datasets. Only few studies in specific datasets showssignificant outcome for both models. As a result, we may deduce that study sparsity and characteristics may have an impact on the results. Despite the fact that the group is imbalanced, the robustness of the research in the dataset can be shown.

Table 3: OR estimate of the dataset D15 with 95% confidence interval estimates

		M1			**M2**		
Dataset	**Title**	**OR Estimate**	**Lower Limit**	**Upper Limit**	**OR Estimate**	**Lower Limit**	**Upper Limit**
D15	Study1	0.30	0.10	0.90	0.37	0.13	1.07
	Study2	0.23	0.05	1.01	0.30	0.07	1.25

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4338-4355
https://publishoa.com
ISSN: 1309-3452

Study3	0.24	0.06	0.74	0.29	0.09	0.92
Study4	0.17	0.05	0.46	0.21	0.06	0.70
Study5	0.26	0.06	0.91	0.32	0.09	1.07
Study6	0.22	0.05	0.86	0.28	0.06	1.16
Study7	0.29	0.15	0.54	0.31	0.17	0.59

Table 3 shows that this dataset contains a total of seven imbalanced studies. Five studies contain no zeros, whereas studies 2 and 6 contain precisely one zero. Study 2, an imbalanced with one zero, yields insignificant outcomes as its confidence interval exceeds one. The remaining studies show that the M1 model produces significant results. Study 1 has the greatest OR 0.303 with interval estimates (0.096, 0.904), while Study 4 has the lowest OR 0.165 with interval estimates (0.047, 0.456). The overall estimate of 0.24 shows that the study has a lower likelihood of favouring the exposed group, with a variance of 40%.

Only studies 3, 4, and 7 are statistically significant for M2 since their OR is lesser than one. The other studies seem to have an OR less than one, but their interval limit is greater than one, thus they are not statistically significant. For M2, studies 1,5 and 6 become insignificant. This fluctuation might be due to the imbalanced nature of the studies. Both M1 and M2 exhibit significant results in studies 3, 4 and 7. As a result, we can see the robustness of these studies.

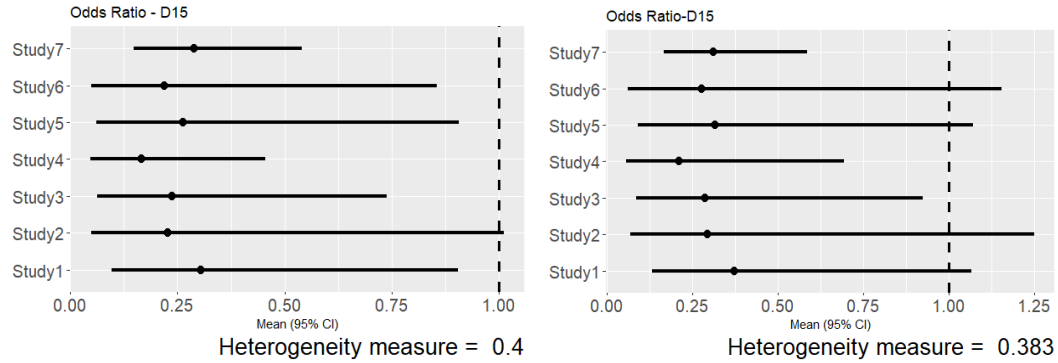


Fig 1 & 2 are the forest plots of D15 for M1 and M2 models.

Only study 2 meets the line of null effect for M1 in the forest plot, yielding an insignificant result. Except for studies 3, 4, and 7, all of the other studies in M2 cross the line of null effect and provide insignificant results. As a result, we may assume that this variation is owing to the sparseness in the studies.

3.2.2 Group II : $10 \leq K < 20$

Eleven datasets are included in this group, where studies in this dataset contains one or more than one zero. Only four datasets shows significant result for both the models.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4338-4355
https://publishoa.com
ISSN: 1309-3452

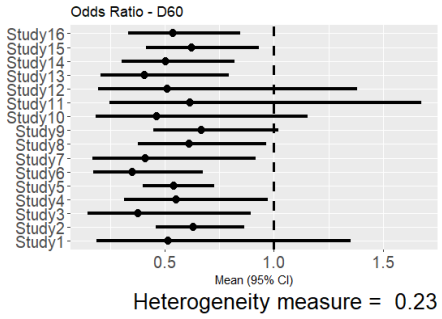
Table 4: OR estimate of the dataset D60 with 95% confidence interval estimates

		M1			M2		
Dataset	**Title**	**OR Estimate**	**Lower Limit**	**Upper Limit**	**OR Estimate**	**Lower Limit**	**Upper Limit**
D60	Study1	0.513	0.185	1.351	0.538	0.206	1.406
	Study2	0.630	0.459	0.865	0.636	0.466	0.868
	Study3	0.375	0.144	0.894	0.499	0.187	1.264
	Study4	0.552	0.312	0.973	0.566	0.325	0.984
	Study5	0.542	0.399	0.726	0.544	0.404	0.736
	Study6	0.352	0.170	0.676	0.380	0.182	0.753
	Study7	0.408	0.168	0.916	0.448	0.176	1.035
	Study8	0.612	0.378	0.966	0.624	0.396	0.980
	Study9	0.668	0.447	1.019	0.680	0.455	1.026
	Study10	0.462	0.181	1.156	0.496	0.199	1.198
	Study11	0.614	0.245	1.675	0.624	0.258	1.574
	Study12	0.510	0.194	1.384	0.540	0.200	1.433
	Study13	0.406	0.206	0.793	0.424	0.216	0.801
	Study14	0.504	0.301	0.819	0.517	0.315	0.842
	Study15	0.620	0.414	0.930	0.625	0.415	0.945
	Study16	0.537	0.332	0.845	0.546	0.339	0.880

From Table 4, it is observed that all the studies seemed to have an OR less than 1, however studies 1,9,10,11, and 12 are considered insignificant since their confidence interval exceeds 1. The outcomes of the remaining 11 studies (2,3,4,5,6,7,8,13,14,15, and 16) are significant. Study 9 has the greatest OR of 0.668 (interval estimate: 0.446, 1.019), while study 6 has the lowest OR of 0.351 (95 % confidence interval: 0.169, 0.675). With a heterogeneity of 0.23, the overall OR is 0.51. According to the results of the present findings, M1 is significant in the majority of the highly skewed studies in this dataset.

In perspective of M2, the outcome of the following seven studies are insignificant: 1,3,7,9,10,11, and 12. The imbalanced group's studies 3 with one zero and 7 with no zero yield insignificant conclusions. This might be related to the study's data qualities. There is a substantial outcome in the remaining trials 2,4,5,6,8,13,14,15, and 16. Study 9 has the highest OR of 0.68, whereas study 6 has the lowest (0.38).

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4338-4355
https://publishoa.com
ISSN: 1309-3452



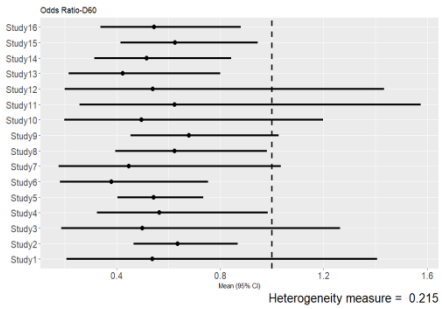


Fig 3 & 4 are the forest plots of D15 for M1 and M2 models.

M2 widens the forest plot of study 3 compared to M1. There was not much of a difference between the models in the remaining studies.

3.2.3. Group III :20 ≤ K <30

Three datasets of ten shows significance in few of its studies. And these studies show significance for both the models M1 and M2.

Table 5: OR estimate of the dataset D13 with 95% confidence interval estimates

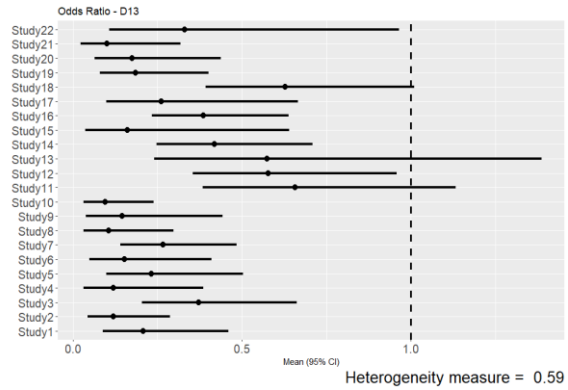
		M1		M2			
Datasets	**Title**	**OR Estimate**	**Lower Limit**	**OR Estimate**	**Lower Limit**	**OR Estimate**	**Lower Limit**
D13	Study1	0.21	0.09	0.46	0.23	0.11	0.50
	Study2	0.12	0.04	0.29	0.15	0.06	0.37
	Study3	0.37	0.20	0.66	0.38	0.22	0.68
	Study4	0.12	0.03	0.39	0.18	0.05	0.56
	Study5	0.23	0.10	0.50	0.25	0.12	0.51
	Study6	0.15	0.05	0.41	0.20	0.07	0.52
	Study7	0.27	0.14	0.48	0.28	0.15	0.52
	Study8	0.11	0.03	0.30	0.17	0.05	0.50
	Study9	0.15	0.04	0.44	0.20	0.06	0.61
	Study10	0.10	0.03	0.24	0.13	0.05	0.34
	Study11	0.66	0.38	1.13	0.65	0.38	1.12
	Study12	0.58	0.35	0.96	0.58	0.36	0.96
	Study13	0.57	0.24	1.39	0.57	0.26	1.33
	Study14	0.42	0.25	0.71	0.43	0.26	0.72
	Study15	0.16	0.04	0.64	0.26	0.06	0.91
	Study16	0.39	0.23	0.64	0.39	0.24	0.63
	Study17	0.26	0.10	0.67	0.29	0.12	0.71
	Study18	0.63	0.39	1.01	0.62	0.39	1.02

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4338-4355
https://publishoa.com
ISSN: 1309-3452

Study19	0.19	0.08	0.40	0.21	0.10	0.45
Study20	0.17	0.06	0.44	0.21	0.08	0.52
Study21	0.10	0.02	0.32	0.24	0.06	0.88
Study22	0.33	0.11	0.96	0.37	0.14	0.95

Table 5 displays the detail of D13. This dataset actually includes 22 studies, with 17 balanced and 5 imbalanced. In M1, though the ORs for studies 11[0.65: (0.38, 1.131), 13[0.57;(0.24,1.38)], and 18[0.62: (0.39, 1.009)] are less than 1, their confidence intervals are greater than one, indicating that their results are insignificant. Study 15 has an OR of 0.16 with a 95% confidence interval of (0.03, 0.63), while Study 21 has an OR of 0.101 with interval estimates of (0.02,0.318). Both studies 15 and 21 belong to the category of unbalanced studies, which have one zero in its cell. Study17, a balanced study with no zero, is statistically significant. The remaining studies 1 to 10, 12, 14, 15, 16, 19, 20, and 22 also show significant results.

For M2, the OR of 15 [0.258: (0.064, 0.914)] and 21 [0.243: (0.058, 0.879)] studies is more than 1. As a result, these studies are regarded as statistically significant. The OR in studies 11[0.653: (0.381, 1.121)], 13[0.57: (0.257, 1.328)], and 18[0.623: (0.389, 1.016)] is less than 1, but the confidence interval bounds are beyond the null effect line, therefore these three studies are not statistically significant. The outcomes of investigations 1 through 10, 12, 14, 15, 16, 19, 20, and 22 are also significant.



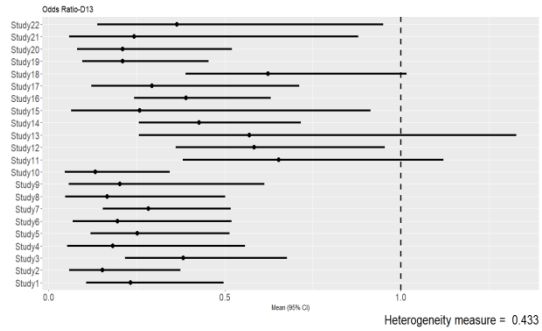


Fig 5 & 6 are the forest plots of D13 for M1 and M2 models.

It has been observed that the forest Plot for a few studies 1 to 10 has become wider in M2 than in M1. Other studies for both M1 and M2 follow a similar pattern. As a result, there was no much variation observed.

3.2.4. Group III : 30 ≤ K < 40

This category includes two datasets, but only D82, shows significant results in a handful of its studies. There are 35 studies in D82, 32 of which are imbalanced. There are no zeros in 29 studies, while there are zeros in 6 studies. From Table 6, we could observe

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4338-4355
https://publishoa.com
ISSN: 1309-3452

that only three of the studies are statistically significant: 1, 6, and 11. Study 1's OR is 0.564 with a 0.05 significance range of (0.359, 0.873), Study 6's OR is 0.703 with a 95 % confidence interval estimate of (0.494, 0.995), and Study 11's OR is 0.649 with a 95 % confidence interval estimate of (0.421, 0.986). Despite the fact that they are imbalanced, these studies are statistically significant. Only two studies 1 and 11 are significant for M2. With a variance of 20%, the overall estimate of 0.95 indicates that the study had a lower likelihood of favouring the exposed group.

Table 6: OR estimate of the dataset D82 with 95% confidence interval estimates

		M1			M2		
Dataset	**Title**	**OR Estimate**	**Lower Limit**	**Upper Limit**	**OR Estimate**	**Lower Limit**	**Upper Limit**
D82	Study1	0.565	0.360	0.873	0.573	0.366	0.885
	Study2	1.133	0.745	1.700	1.130	0.746	1.720
	Study3	0.694	0.433	1.100	0.700	0.447	1.096
	Study4	0.729	0.312	1.605	0.744	0.323	1.657
	Study5	0.756	0.304	1.711	0.909	0.372	2.167
	Study6	0.703	0.495	0.995	0.709	0.500	1.000
	Study7	0.741	0.490	1.105	0.740	0.491	1.108
	Study8	0.889	0.521	1.516	0.891	0.525	1.490
	Study9	0.973	0.755	1.253	0.975	0.745	1.266
	Study10	1.014	0.679	1.548	0.999	0.661	1.507
	Study11	0.650	0.422	0.986	0.657	0.434	0.994
	Study12	1.247	0.683	2.440	1.240	0.682	2.274
	Study13	1.138	0.624	2.132	1.144	0.647	2.048
	Study14	0.756	0.310	1.692	0.912	0.372	2.185
	Study15	1.006	0.905	1.118	1.007	0.907	1.117
	Study16	0.953	0.731	1.228	0.954	0.735	1.240
	Study17	0.885	0.549	1.406	0.891	0.562	1.413
	Study18	0.969	0.650	1.449	0.974	0.654	1.457
	Study19	0.896	0.420	1.943	0.907	0.434	1.886
	Study20	0.789	0.525	1.164	0.792	0.535	1.173
	Study21	1.393	0.886	2.190	1.380	0.879	2.162
	Study22	1.287	0.597	2.928	1.182	0.523	2.757
	Study23	0.859	0.415	1.805	0.868	0.429	1.743
	Study24	1.076	0.911	1.279	1.077	0.906	1.285
	Study25	0.782	0.536	1.126	0.790	0.550	1.127
	Study26	0.874	0.347	2.028	0.929	0.384	2.219

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4338-4355
https://publishoa.com
ISSN: 1309-3452

Study27	1.305	0.753	2.360	1.300	0.749	2.285
Study28	0.949	0.698	1.298	0.955	0.694	1.320
Study29	0.999	0.675	1.455	1.004	0.685	1.469
Study30	1.496	0.791	2.959	1.373	0.690	2.866
Study31	2.455	1.957	3.132	2.436	1.915	3.095
Study32	1.036	0.433	2.471	1.000	0.418	2.394
Study33	0.860	0.343	2.013	0.927	0.389	2.190
Study34	1.024	0.445	2.496	0.994	0.413	2.396
Study35	0.891	0.378	2.091	0.917	0.405	2.080

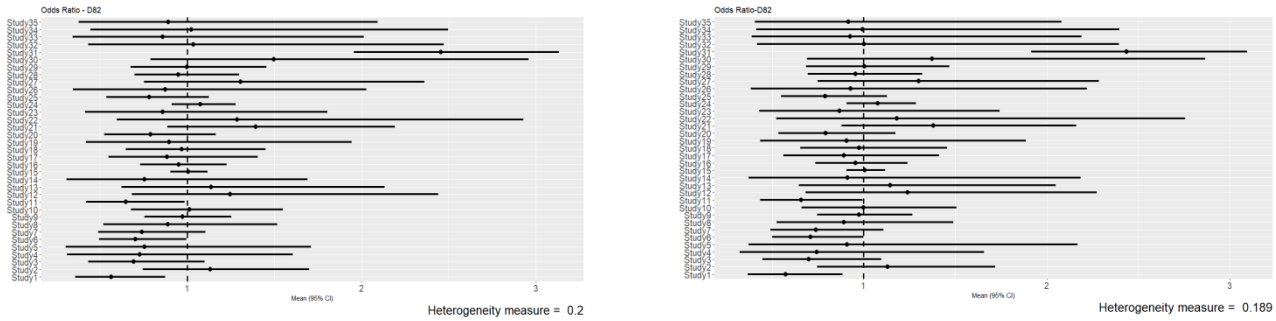


Fig 7 & 8 are the forest plots of D82 for M1 and M2 models.

Study 6 yields an insignificant result since its confidence interval crosses the null effect line. As a result, we could see the robustness studies 1 and 11 from the forest plot. The remaining studies for M1 and M2 follow a similar structure.

3.2.5. Group III : 40 ≤ K < 50

This category contains four datasets. Only one dataset, D6, yields a significant outcome, whereas the other three are insignificant. There are forty-one studies in Efron's 1996 dataset. D6 is a dataset with seven balanced studies and thirty-four imbalanced studies in it. This dataset contains thirty-one studies with no zeros, six studies with one zero, three studies with two zeros, and Study 4 is an unique case in which all of the arms contain zero.

Table 7: OR estimate of the dataset D6 with 95% confidence interval estimates

		M1			M2		
Dataset	**Title**	**OR Estimate**	**Lower Limit**	**Upper limit**	**OR Estimate**	**Lower Limit**	**Upper limit**
D6	Study1	0.15	0.03	0.67	0.27	0.08	0.83
	Study2	0.56	0.17	1.96	0.52	0.20	1.37

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4338-4355
https://publishoa.com
ISSN: 1309-3452

Study3	1.06	0.28	4.07	0.72	0.27	2.07
Study4	0.19	0.01	4.71	0.37	0.08	1.68
Study5	2.55	0.31	29.82	0.45	0.11	2.03
Study6	0.08	0.00	0.93	0.32	0.07	1.29
Study7	0.23	0.06	0.82	0.31	0.11	0.84
Study8	0.03	0.00	0.15	0.16	0.04	0.56
Study9	0.42	0.09	1.79	0.44	0.16	1.29
Study10	0.10	0.02	0.34	0.20	0.06	0.57
Study11	0.05	0.00	0.48	0.31	0.07	1.24
Study12	0.12	0.02	0.72	0.27	0.07	0.88
Study13	1.35	0.42	4.52	0.87	0.35	2.39
Study14	1.41	0.40	5.20	0.86	0.32	2.48
Study15	0.22	0.05	0.81	0.31	0.11	0.84
Study16	0.30	0.05	1.47	0.37	0.12	1.08
Study17	0.08	0.01	0.42	0.22	0.06	0.71
Study18	0.17	0.02	1.02	0.30	0.09	1.03
Study19	0.17	0.05	0.54	0.25	0.09	0.65
Study20	0.30	0.11	0.80	0.34	0.15	0.77
Study21	0.11	0.03	0.35	0.19	0.07	0.48
Study22	1.21	0.28	5.08	0.79	0.29	2.39
Study23	1.07	0.46	2.42	0.86	0.41	1.83
Study24	0.54	0.17	1.82	0.51	0.20	1.35
Study25	0.02	0.00	0.18	0.30	0.06	1.26
Study26	0.18	0.02	1.15	0.32	0.09	1.06
Study27	0.62	0.15	2.38	0.53	0.19	1.54
Study28	0.02	0.00	0.12	0.28	0.06	1.16
Study29	0.04	0.00	0.30	0.30	0.06	1.21
Study30	0.13	0.03	0.53	0.25	0.08	0.74
Study31	0.08	0.01	0.41	0.22	0.06	0.74
Study32	0.40	0.10	1.42	0.41	0.16	1.11
Study33	0.10	0.02	0.52	0.24	0.07	0.75
Study34	0.04	0.00	0.33	0.30	0.07	1.23
Study35	1.18	0.24	6.67	0.68	0.23	2.29
Study36	0.21	0.07	0.60	0.27	0.11	0.66
Study37	0.88	0.15	4.90	0.60	0.19	2.02
Study38	0.32	0.08	1.18	0.37	0.13	1.00
Study39	0.30	0.11	0.82	0.34	0.15	0.79
Study40	0.00	0.00	0.01	0.28	0.06	1.11
Study 41	0.19	0.01	5.09	0.38	0.09	1.64

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4338-4355
https://publishoa.com
ISSN: 1309-3452

From Table 7, it is observed that twenty-two studies show significant result. The remaining studies are not statistically significant. The overall OR estimate is 0.19 with heterogeneity measure of 2.68 for M1.

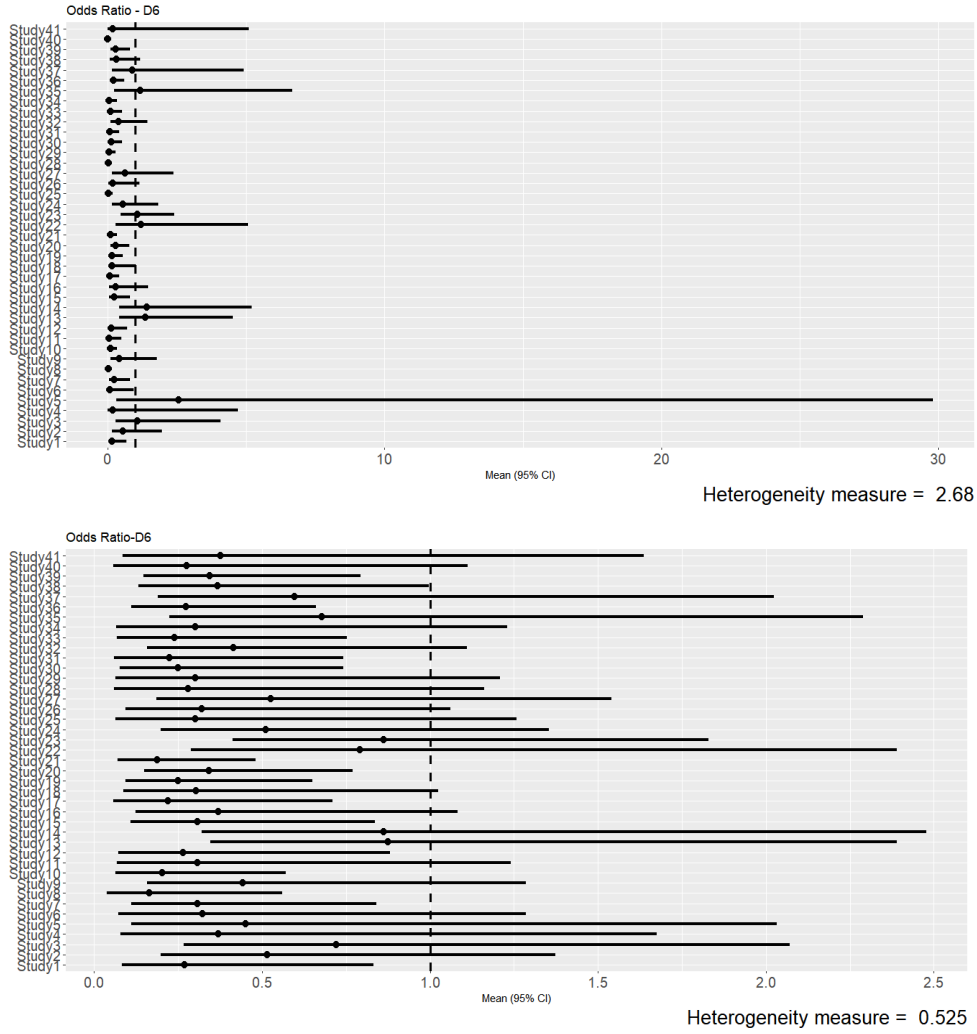


Fig 9&10 are the forest plots of D6 for M1 and M2 models.

Only 15 studies show a significant outcome for M2. For M2, the remaining studies 6, 11, 25, 28, 29, 34, 38, and 40 become insignificant. As a result, the aforementioned 15 studies demonstrate its reliability for both the M1 and M2 models. The majority of studies in D6 achieved significant results for both M1 and M2. As a consequence, we may conclude that study sparsity and characteristics may influence the outcomes. Despite the fact that the group is imbalanced, the dataset's robustness may be established.

4. Summary

The data used in analysis consisted of 83 datasets that have been used to look at treatment variation among datasets. Since the cell frequencies in certain studies were so minimal, the studies were sparse. In such situations, the Bayesian technique appears to be more flexible since it is independent of the

study design and allows for the integration of data from several sources, as is required in clinical studies. However, like any Bayesian analysis, establishing priors that are compatible with prior knowledge of the issue and data is critical, and the implementation process takes on added importance.

The present study has demonstrated that when pooling several studies with sparse data, it is fundamental to use the Binomial – Normal and Normal-Normal models. In the presence of heterogeneity, however, study-specific data must be assessed in depth in addition to overall estimates. The primary goal is to make reasonable conclusions about the overall impact measure and between-variance. As a result, the goal of this paper is to give a thorough list of methods for meta-analytic techniques, with a focus on between-variance estimators under REM and their influence on point and interval dataset estimates. A hierarchical Bayesian model was used to illustrate individual and overall odds ratios for sparse datasets.

The analysis section demonstrates how well highly imbalanced sparse studies perform in Binomial-Normal and Normal-Normal models for datasets with varied k values. In group I, study in more than 60% of the dataset reveals significant findings. The second group (II), which includes 10 to 20 research, offers a dataset. In this section, significant aspects are identified from 36% of the dataset.30 percent of the dataset in group III is significant. Sections IV and V, meanwhile, provide significant results in 50% and 25% of the datasets, respectively.

We could see from the above analysis that as the number of studies increases, there is a variance in the dataset's outcomes. The dataset'sconsistency, however, has been noticed. For all datasets, a positive heterogeneity measure is obtained, indicating that there is a variance between the studies in the datasets. Both approaches outperformed in all other sparse datasets and are recommended for use in comparable meta-analyses. Finally, the model to adopt is related to the number of studies and data nature.

As a consequence, the study includes a summary of the findings as well as an evaluation of various clinical data components. The current study is not complete in and of itself, and there is always scope for improvements; for the sake of research continuation, a few prospective expansions have been highlighted. The discussion has been structured thus far in the process of dealing with sparse data in a Bayesian framework.

REFERENCES

1. Abrams, K., &Sansó, B. (1998). Approximate Bayesian inference for random effects meta-analysis. Statistics in medicine, 17(2), 201-218.
2. Agresti, A. (2013). Categorical Data Analysis. Wiley.
3. Agresti, A., & Hartzel, J. (2000). Strategies for comparing treatments on a binary response with multi-centre data. Statistics in medicine, 19(8), 1115-1139.
4. Bowden, J., Tierney, J. F., Copas, A. J., and Burdett, S. (2011). Quantifying, displaying and accounting for heterogeneity in the meta-analysis of rcts using standard

and generalised q statistics. BMC medical research methodology, 11(1):41.

5. Card, N. A. (2012). Applied meta-analysis for social science research. Guilford Publications. New York
6. Carlin, J. B. (1992). Meta-analysis for 2× 2 tables: a Bayesian approach. Statistics in medicine, 11(2), 141-158.
7. Chu, Y., & Yuan, Y. (2018). A Bayesian basket trial design using a calibrated Bayesian hierarchical model. Clinical Trials, 15(2), 149-158.
8. Davis, J., Mengersen, K., Bennett, S., and Mazerolle, L. (2014). Viewing systematic reviews and meta-analysis in social research through different lenses. SpringerPlus, 3(1):511.
9. Efron, B. (1996). Empirical Bayes methods for combining likelihoods. Journal of the American Statistical Association, 91(434), 538-550.
10. Engels, E. A., Schmid, C. H., Terrin, N., Olkin, I., and Lau, J. (2000). Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. Statistics in medicine, 19(13):1707–1728.
11. Fleiss, J. L. (1986). Analysis of data from multiclinic trials. Controlled clinical trials, 7(4), 267-275.
12. Friedman, L. M., Furberg, C. D., DeMets, D. L., Reboussin, D. M., & Granger, C. B. (2015). Fundamentals of clinical trials. springer.
13. Frison, L., & Pocock, S. J. (1992). Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design. Statistics in medicine, 11(13), 1685-1704.
14. Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian data analysis. CRC press.
15. Green, P. E., & Park, T. (2004). Bayesian methods for contingency tables using Gibbs sampling. Statistical Papers, 45(1), 33-50.
16. Higgins, J. and Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. Statistics in medicine, 21(11):1539–1558
17. Hobbs, B. P., & Carlin, B. P. (2007). Practical Bayesian design and analysis for drug and device clinical trials. Journal of biopharmaceutical statistics, 18(1), 54-80.
18. Johnson, J. R., Williams, G., &Pazdur, R. (2003). End points and United States Food and Drug Administration approval of oncology drugs. Journal of Clinical Oncology, 21(7), 1404-1411.
19. Langan, D., Higgins, J., and Simmonds, M. (2016). Comparative performance of heterogeneity variance estimators in meta-analysis: a review of simulation studies. Research synthesis methods
20. Lewis, S. and Clarke, M. (2001). Forest plots: trying to see the wood and the trees. BMJ: British Medical Journal, 322(7300):1479.
21. Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. BMC medical informatics and decision making, 11(1), 1-13.

22. Piantadosi, S. (2017). Clinical trials: a methodologic perspective. John Wiley & Sons.
23. Pezeshk, H., & Gittins, J. (2002). A fully Bayesian approach to calculating sample sizes for clinical trials with binary responses. Drug Information Journal, 36(1), 143-150.
24. R Core Team (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
25. Riley, R. D., Higgins, J. P., and Deeks, J. J. (2011). Interpretation of random effects meta-analyses. Bmj, 342:d549.
26. Senn, S. (2001). Cross-over trials in drug development: theory and practice. Journal of statistical planning and inference, 96(1), 29-40.
27. Skene, A. M., & Wakefield, J. C. (1990). Hierarchical models for multicentre binary response studies. Statistics in Medicine, 9(8), 919-929.
28. Smith, J. A. (1995). Semi structured interviewing and qualitative analysis.
29. Sorensen, D., Gianola, D., &Gianola, D. (2002). Likelihood, Bayesian and MCMC methods in quantitative genetics.
30. Subbiah, M., & Srinivasan, M. R. (2008). Classification of 2× 2 sparse data sets with zero cells. Statistics & probability letters, 78(18), 3212-3215.
31. Tian, L., Cai, T., Piankov, N., Cremieux, P. Y., & Wei, L. J. (2007). 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.
32. Viechtbauer, W. (2007). Confidence intervals for the amount of heterogeneity in metaanalysis. Statistics in medicine, 26(1):37–52.
33. Viechtbauer, W. (2010). Conducting meta-analyses in r with the metafor package. Journal of Statistical Software, 36(3):1–48
34. Wu, Y., Tjelmeland, H., & West, M. (2007). Bayesian CART: Prior specification and posterior simulation. Journal of Computational and Graphical Statistics, 16(1), 44-66.
35. Yin, G., Li, Y., & Ji, Y. (2006). Bayesian dose-finding in phase I/II clinical trials using toxicity and efficacy odds ratios. Biometrics, 62(3), 777-787.