A Comparison of Some Methods for Detection of Safety Signals in Randomised Controlled Trials

Raymond Carragher
University of Strathclyde
raymond.carragher@strath.ac.uk

Abstract

The occurrence, severity and duration of patient adverse events are routinely recorded during randomised controlled clinical trials. This data may be used by a trial's Data Safety Monitoring Committee to make decisions regarding the safety of treatments and in some cases may lead to the discontinuation of a trial if real safety issues are detected. Consequently the analysis of this data is a very important part of the conduct of any trial.

There are many different types of adverse event and the statistical analysis of this data must take into account multiple comparison issues when performing statistical tests. Unadjusted tests may lead to large numbers of false positive results, but simple adjustments are generally too conservative and risk compromising the power to detect important treatment differences. Mathematically there are a number of different approaches to analysing safety data with general error controlling procedures, recurrent event analysis, survival analysis and other direct modelling approaches (both Bayesian and Frequentist) all being used. Recently a variety of classical (Mehrotra and Adewale, 2012) and Bayesian (Berry and Berry, 2004; DuMouchel, 2010) methods have been proposed to address this problem. These methods use possible relationships or groupings of the adverse events.

We implement and compare by way of a simulation study of grouped data some of these more recent approaches to adverse event analysis and investigate if the use of a common underlying model which involves groupings of adverse events by body-system or System Organ Class is useful in detecting adverse events associated with treatments. All of the group methods detect more correct significant effects than the Benjamini-Hochberg or Bonferroni procedures for this type of data. In particular the body-system as described by Berry and Berry (2004) looks to be a worthwhile structure to consider for use when modelling adverse event data.

1 Introduction

Randomised controlled clinical trials, conducted under the supervision of a Data (Safety) Monitoring Committee (DMC), are the standard method for establishing the efficacy and safety of new treatments. During a trial the DMC may, based on the analysis of clinical events recorded during the trial, make recommendations about the conduct of the trial. This can include recommendations regarding the termination of the whole trial or individual trial "arms". Reasons for terminations before the scheduled end of the trial can include early demonstrations of efficacy, concerns regarding safety issues, or the possibility that continuing the trial may be futile. Consequently the analysis of the safety related data, in particular what are termed Adverse Events (AEs), is extremely important.

The anticipated effect sizes of Adverse Events in these studies are generally small and in order to accumulate the number of events to detect such effect sizes with a sufficiently high power the follow up time has to be very long or a large number of patients need to be recruited. Due to the large number of different variables recorded unadjusted significance tests may lead to large numbers of false positive results, but simple adjustments risk compromising the already possibly low power to detect important treatment differences. Recently a variety of classical ([1]) and Bayesian ([2], [3]) methods have been proposed to address this problem. These papers consider how the use of relationships which may exist between Adverse Events may be used to group them into body-systems and this additional information used in a statistical analysis. In a Bayesian context, the information within the body-systems can also be used as one approach to handling multiplicities with the additional information available used to shrink non-significant effects towards zero and to borrow strength from the assumed relationships among the Adverse Events ([4]).

These methods are relatively complex to implement and there is to date little experience among practitioners in their use. In this study we implemented a number of the methods to compare their interpretability and error rates on simulated data and to investigate if the body-system is useful when dealing with safety data. For the simulation we used 8 different body-systems, numbered 1 - 8, with between 1 and 11 different Adverse Events in each body-system. There were 45 Adverse Events in total.

2 Methods and Implementation

The following methods for analysing Adverse Event incidence data are compared: unadjusted significance testing (NOADJ); the Bonferroni correction (BONF) ([5]); control of the False Discovery Rate (FDR) by the Benjamini-Hochberg procedure ([6]); control of the False Discovery Rate by a number of grouping procedures: the Double False Discovery Rate (DFDR) ([1]), Group Benjamini-Hochberg (GBH) ([7]), subset Benjamini-Hochberg (ssBH) ([8]; the three-level hierarchical model of Berry and Berry ([2]) and Model 1a from Xia et al ([9]) which is a subset of the Berry and Berry model, details of which are given in §A.2, §A.3.

The unadjusted testing, Bonferroni correction and False Discovery Rate control procedures all require the calculation of p-values. For the purposes of the simulation we followed [2] and use an exact Fisher 2-sided test. Direct comparisons between these error controlling procedures are possible. However direct comparisons with the Bayesian models require that the Bayesian approaches have a defined criteria for flagging Adverse Events. None of the methods we are looking at have such definitive criteria so when comparing across the different methods we used a nominal threshold value of 95% posterior probability for the Bayesian methods. For the grouped BH methods (DFDR, GBH, ssBH) we used threshold values of 5%. The ssBH method uses groupings of hypotheses to extend the range of dependent test statistics to which a BH type FDR controlling procedure can be applied and still control the FDR at the desired level.

It is known to be as or less powerful than the BH-procedure itself in all circumstances.

The methods that are studied are:

Method Name	**Description**
HIER.BB	Berry and Berry model, [2], also model 1b from [9]
HIER.1a	Model 1a from [9]
BH	False discover rate control by the Benjamini-Hochberg procedure ([6])
DFDR	Double false discovery rate ([1])
NOADJ	Unadjusted significance testing
BONF	Bonferroni correction ([5])
GBH	Group Benjamini-Hochberg ([7])
ssBH	Subset Benjamini-Hochberg ([8])

Table 1: Methods used in Simulation Study

All the methods were implemented as R-packages (§A.1). The Bayesian models (HIER.1a, HIER.BB) were fitted using MCMC methods.

3 Simulation Study

Simulated Body-System Trial Data

The simulated data for Adverse Events based on the body-systems used a logistic regression model to generate the trial incidence data. The data is marginal and assumed to correspond to the binomial model (following [2]):

$$\begin{array}{ll} \text{Controls:} & X_{bj} = \text{Bin}(N_C, c_{bj}) \\ \text{Treatments:} & Y_{bj} = \text{Bin}(N_T, t_{bj}) \end{array} \tag{1}$$

The data model in its most general form is:

$$\text{logit}(p_{tbj}) = \mu_{tbj} + U_{tbj} \tag{2}$$

where

- $t = 1, 2$ corresponding to the control and treatment groups respectively and p_{tbj} is the probability of the occurrence of the j^{th} Adverse Event in body system b in treatment group t. We have $c_{bj} = p_{1bj}$ and $t_{bj} = p_{2bj}$ as per [2].
- μ_{tbj} and U_{tbj} are fixed and random underlying rates for Adverse Event j in body-system b and treatment group t respectively.

For the purposes of the simulation we are interested in detecting increases in the odds ratios or relative risks of Adverse Events between the two groups.

We used 8 body-system groupings, numbered 1 - 8, containing 1, 4, 7, 5, 9, 11, 3 and 6 different Adverse Events respectively and we looked at three separate trial sizes: Trial 1 ($N_C = 110, N_T = 110$), Trial 2 ($N_C = 450, N_T = 450$) and Trial 3 ($N_C = 1100, N_T = 1100$).

Simulation Results

In this section we look at the results from one particular repeated simulation where the underlying Adverse Event rate is raised for body system 5 for both treatment and control, raised for body-system 3 for treatment only and raised for Adverse Events 1, 2 in body-system 2 for

treatment only. The plots below (Figure 1) shows how the estimated parameters match the actual parameters underlying the model, TDM15, without the random effects. For both control and treatment the models have successfully estimated the parameter values.

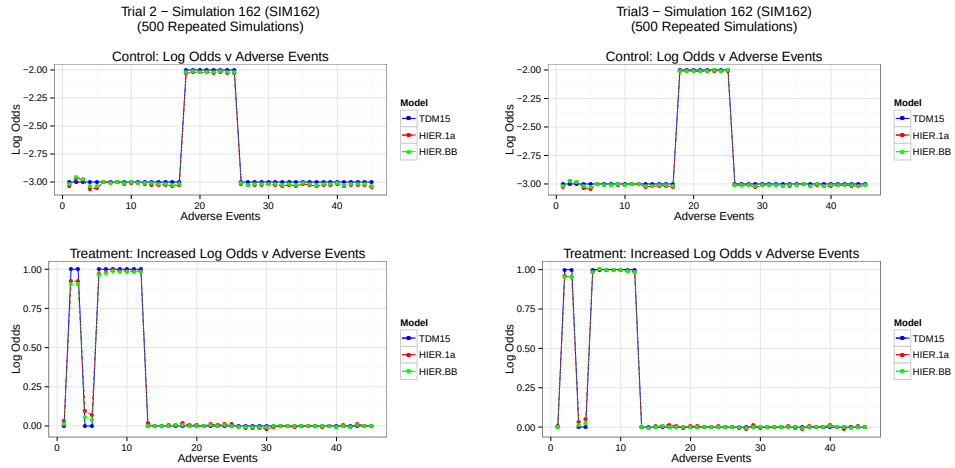

Figure 1: Estimated Log Odds and Increased Log Odds (Parameter Estimates)

The following plots (Figure 2) show the posterior distributions for the single Adverse Event in body system 1 and the first Adverse Event in body system 3 for one simulation.

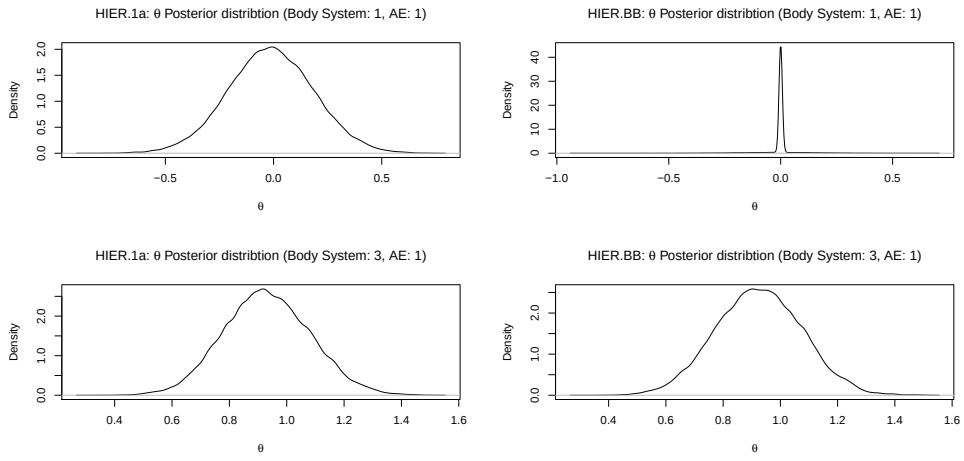

Figure 2: Posterior Distributions

For body-system 1, where there is no difference between treatment and control, the distribution for the no-point-mass model (HIER.1a) is centred on 0 but has a relatively large spread. On the other hand for the point-mass model (HIER.BB) the posterior distribution is also effectively a point-mass at 0. For body system 3 both posteriors are similar and the distribution is almost totally greater than zero (as expected).

The results for each model for the repeated simulation are as follows:

Model	Correct	Incorrect	Missed
HIER.BB	4303	9	197
HIER.1a	4492	582	8
NOADJ	4374	682	126
BONF	3258	12	1242
DFDR	4317	72	183
BH	4022	114	478
GBH	4441	144	59
ssBH	3848	14	652

Table 2: Trial 2

Model	Correct	Incorrect	Missed
HIER.BB	4498	5	2
HIER.1a	4500	705	0
NOADJ	4500	707	0
BONF	4486	10	14
DFDR	4500	67	0
BH	4499	132	1
GBH	4500	143	0
ssBH	4498	25	2

Table 3: Trial 3

where Correct are the counts of raised rate Adverse Events detected (an estimate of the power), Incorrect are counts of Adverse Events whose rate is not raised who are declared significant (Type-I errors) and Missed are are the counts of raised rate Adverse Events which were not detected (essentially an estimate of the Type-II error rate).

We can see that for the largest trial (Trial 3) most of the methods correctly identify the Adverse Events with raised rates and the numbers missed are low. However some of the methods have a tendency to incorrectly identify Adverse Events. The model HIER.BB appears to perform best overall. As might be expected the results are not as clearcut for Trial 2. In this case the method, HIER.1a, which indentifies the most Adverse Events also has the second highest incorrectly identification count but the lowest Type-II error rate. Again, an argument could be made that method HIER.BB performs best overall. The results for the simulations for Trial 1 were similar but more variable. It should be noted that the cut-off for the Bayesian models, HIER.1a, HIER.BB, was an arbitrarily chosen 95%. With a different cut-off value the results would have been different.

The simulation discussed in this section is for body-system simulated data. Similar simulations based on independent Adverse Event data simulation models have been done and show less discrepancies between the models.

4 Discussion

There is a need to be careful when drawing conclusions from simulation studies. The cut-off chosen for the Bayesian models, HIER.1a and HIER.BB, was somewhat arbitrary at 95%. A different choice of cut-off for each model could have been used to gain better results, increasing the power without inflating the Type-I error. For our purposes the use of a 95% cut-off allows suitable comparisons to be made between the methods and the determination of such a cut-off is not the main point of the simulation study at this stage.

The main conclusion from the simulations is that, for data where there are believed to be relationships between the Adverse Events, using groupings (body-systems) does appear to make a difference to the results. All of the group methods, with the exception of ssBH as noted above, detect more correct significant effects than the Benjamini-Hochberg or Bonferroni procedures. However, for some of these methods the Type-I error may become inflated in comparison to the other methods, e.g. HIER.1a. In particular the body-system as described by [2] and [9] looks to be a worthwhile structure to consider for use when modelling data.

As stated in [2] the point mass in HIER.BB makes a quantitive difference in the modelling approaches. We have seen in the simulations that with the same cut-off point for both models HIER.1a and HIER.BB the effect of the point mass is to both reduce the numbers of correctly detected Adverse Events and also the Type-I error rates. There is a trade-off to be made.

A Models and Implementation Details

A.1 Software

R software:

Model	R-package
HIER.1a	c212.BBa
HIER.BB	c212.BBb
c212.BONF	c212.FDR
c212.DFDR	c212.FDR
c212.FDR	c212.FDR
c212.GBH	c212.FDR
c212.NOADJ	c212.FDR
c212.ssBH	c212.FDR

Table 4: R Software

The software is available at `http://personal.strath.ac.uk/raymond.carragher/`.

A.2 Model HIER.BB

In this model we consider B body systems. Within body system b there are k_b types of Adverse Event labelled AE_{bj} where $b = 1, \ldots, k_b$. There are N_C patients in the control group and N_T in the treatment group.

X_{bj} and Y_{bj} are the number of occurrences of A_{bj} in the control and treatment groups respectively with the probabilities of experiencing A_{bj} being c_{bj} for the control group and t_{bj} for the treatment group. In the following $b = 1 \ldots B$, $j = 1 \ldots k_b$.

The model for the incidence data is:

$$\begin{aligned} X_{bj} &\sim \text{Bin}(N_C, c_{bj}) \\ Y_{bj} &\sim \text{Bin}(N_T, t_{bj}) \end{aligned} \tag{3}$$

where N_C, N_T are the number in the control and treatment groups respectively, X_{bj} is the count of Adverse Event incidence in the control group for the j^{th} Adverse Event in body-system b and Y_{bj} is the count for the treatment group.

The top level of the hierarchy models the log-odds by:

$$\begin{aligned} &\gamma_{bj} = N(\mu_{\gamma b}, \sigma^2_{\gamma b}) \\ &\theta_{bj} = \pi_b I_{[0]} + (1 - \pi_b) N(\mu_{\theta b}, \sigma^2_{\theta b}) \quad b = 1, \ldots, B, \;\; j = 1, \ldots, k_b \end{aligned} \tag{4}$$

where I is the indicator function.

As per usual in hierarchical models each of the parameters $\{\mu_{\theta b}, \mu_{\gamma b}, ...\}$ are considered to be random variables with their own distributions. The model is fully described in [2].

A.3 Model HIER.1a

Model 1a in [9] is an implementation of the Berry and Berry model ([2]) without the point mass and is exactly the same as Model 1b above apart from the point mass and its corresponding hyperparameters.

References

[1] D. V. Mehrotra and A. J. Adewale, "Flagging clinical adverse experiences: reducing false discoveries without materially compromising power for detecting true signals," *Stat Med*, vol. 31, no. 18, pp. 1918–30, 2012. Mehrotra, Devan V Adewale, Adeniyi J Comparative Study England Stat Med. 2012 Aug 15;31(18):1918-30. doi: 10.1002/sim.5310. Epub 2012 Mar 13.

[2] S. M. Berry and D. A. Berry, "Accounting for multiplicities in assessing drug safety: A three-level hierarchical mixture model," *Biometrics*, vol. 60, no. 2, pp. 418–426, 2004.

[3] W. DuMouchel, "Multivariate bayesian logistic regression for analysis of clinical study safety issues," *Statistical Science*, vol. 27, no. 3, pp. 319–339, 2012.

[4] D. B. Dunson, A. H. Herring, and S. M. Engel, "Bayesian selection and clustering of polymorphisms in functionally related genes," *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 534–546, 2008.

[5] J. N. S. Matthews, *Introduction to Randomized Controlled Clinical Trials, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science, Chapman and Hall/CRC, 2006. doi:10.1201/9781420011302.fmatt.

[6] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[7] J. X. Hu, H. Zhao, and H. H. Zhou, "False discovery rate control with groups," *J Am Stat Assoc*, vol. 105, no. 491, pp. 1215–1227, 2010. Hu, James X Zhao, Hongyu Zhou, Harrison H R01 GM059507-09/GM/NIGMS NIH HHS/United States J Am Stat Assoc. 2010 Sep 1;105(491):1215-1227.

[8] D. Yekutieli, "False discovery rate control for non-positively regression dependent test statistics," *Journal of Statistical Planning and Inference*, vol. 138, no. 2, pp. 405–415, 2008.

[9] H. Amy Xia, H. Ma, and B. P. Carlin, "Bayesian hierarchical modeling for detecting safety signals in clinical trials," *Journal of Biopharmaceutical Statistics*, vol. 21, no. 5, pp. 1006–1029, 2011.