Received: 6 January 2021 | Accepted: 14 June 2021

DOI: 10.1002/pst.2148

MAIN PAPER

WILEY

Assessing safety at the end of clinical trials using system organ classes: A case and comparative study

Raymond Carragher[1,2,3] | **Chris Robertson**[2,4]

[1]Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, Glasgow, UK

[2]Department of Mathematics and Statistics, University of Strathclyde, Glasgow, UK

[3]Health Data Research (UK), University of Strathclyde, Glasgow, UK

[4]Public Health Scotland, NHS National Services Scotland, Glasgow, UK

Correspondence
Raymond Carragher, Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, 161 Cathedral Street, Glasgow, Lanarkshire G4 0RE, UK.
Email: raymond.carragher@strath.ac.uk

Funding information
Engineering and Physical Sciences Research Council, Grant/Award Number: 1521741; UK Research and Innovation, Grant/Award Number: MR/S003967/1

Abstract

Recent approaches to the statistical analysis of adverse event (AE) data in clinical trials have proposed the use of groupings of related AEs, such as by system organ class (SOC). These methods have opened up the possibility of scanning large numbers of AEs while controlling for multiple comparisons, making the comparative performance of the different methods in terms of AE detection and error rates of interest to investigators. We apply two Bayesian models and two procedures for controlling the false discovery rate (FDR), which use groupings of AEs, to real clinical trial safety data. We find that while the Bayesian models are appropriate for the full data set, the error controlling methods only give similar results to the Bayesian methods when low incidence AEs are removed. A simulation study is used to compare the relative performances of the methods. We investigate the differences between the methods over full trial data sets, and over data sets with low incidence AEs and SOCs removed. We find that while the removal of low incidence AEs increases the power of the error controlling procedures, the estimated power of the Bayesian methods remains relatively constant over all data sizes. Automatic removal of low-incidence AEs however does have an effect on the error rates of all the methods, and a clinically guided approach to their removal is needed. Overall we found that the Bayesian approaches are particularly useful for scanning the large amounts of AE data gathered.

KEYWORDS

adverse events, Bayesian hierarchy, false discovery rate, safety, system organ class

1 | INTRODUCTION

The challenges of performing statistical analyses on adverse event (AE) data are well known, with low event rates, low power, small effect sizes and multiple comparison issues all complicating the analysis. This has stimulated the development of methods for determining robust safety profiles for treatments during the trial process, while addressing the associated statistical difficulties. A number of recent approaches, which use relationships in the data to group the AEs into system organ classes (SOCs), have been developed, with both direct error controlling procedures[1,2] (methods which control error rates at a specific level when testing multiple hypotheses) and Bayesian models[3–5] being proposed as

viable methods for dealing with these issues. The classification of AEs into SOCs is typically accomplished using a medical dictionary such as MedDRA,* guided by clinical input.[6]

By using relationships between outcomes grouped methods offer the possibility of detecting more AEs in the premarketing phase of a treatment lifecycle, and with their routine use becoming more common,[7–10] aided by the existence of implementations of a number of the methods,[11,12] an understanding of their suitability is an important factor when analysing the methods' results. Characteristics which may influence the choice of method are not only the power and error rates, but also the effect the data structure has on the methods' performances. While SOCs provide a collection of medically related outcomes, and reporting of AE incidence by SOC is prevalent in clinical trial study reports, they are quite broad in scope, with many of the large numbers of AEs recorded during the trial process having low incidence. While the removal of these low incidence AEs may be a legitimate part of the analysis,[13] the effect that this may have on the methods is largely unexplored. An additional consideration is that while the approaches are all designed to flag AEs as being potentially associated with treatment they differ considerably in their philosophies. The Bayesian approaches generally employ a system of information sharing using a hierarchical structure of AEs and SOCs. The relationship between AEs within a SOC is an important part of the model. In a Bayesian context the information within the SOCs 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 relationships among the AEs.[14,15] In technical terms the models have a common underlying mean SOC level effect.[3] This tacitly assumes a positive relationship between the AEs within a SOC. In the case of more complicated relationships, for example if the AEs are negatively correlated, we might expect there to be large variability in the posterior distribution of the mean SOC level effect, and this will affect how the models determine which AEs are associated with treatment. While these more complicated cases are not specifically investigated, we do investigate how well the models cope with large variations in the AE counts within the SOCs. Bayesian grouped models of this type have also been extended for trial meta-analysis[16] and for observational data.[17] The error controlling approaches are more general, designed to control error rates at a particular level when performing multiple hypothesis testing. The choice of hypothesis test is left to the user with the procedures acting directly on the tests' p-values. While they may employ a grouped structure of the AEs into SOCs,[1,2] and use this in their testing procedure, in a more general sense information sharing is part of the process only in so much as it is part of the underlying hypothesis tests. The methods included in this study are given in Table 1.

The methods are first applied to real trial safety data. Here we find that while the Bayesian models appear to be appropriate for the full data set, the error controlling methods only give similar results to the Bayesian methods when low incidence AEs are removed. This naturally leads to the question of how removing (low incidence) AEs from the analysis affects the methods. A simulation study is used to directly compare the methods both on full data sets and data sets where low incidence AEs are systematically removed. We find that the removal of low incidences AEs increases both the power and the FDR for the error controlling procedures but does not have a major effect on the power of the Bayesian models. However, due to way the Bayesian models' information sharing mechanism works, the FDR of the Bayesian methods may vary on the removal of low incidence AEs from SOCs which are still included in the analysis. The removal of complete SOCs in the analysis does not have the same effect on the grouped methods, indicating that within SOC relationships are a stronger factor when determining which AEs are associated with treatment. We further investigate the effect of increasing the numbers of AEs with changed treatment occurrence rates within SOC. This is potentially important as AEs with increased rates but which themselves may not be of clinical interest have an effect on the results, and we find that overall the detection rates of the AEs are increased.

TABLE 1 Methods included in the study

Method name	AE data grouped by SOC	Direct error control	Description
BB	Yes	None (Bayesian model)	Berry and Berry model[3,5]
1a	Yes	None (Bayesian model)	Berry and Berry model without point-mass[5]
DFDR	Yes	FDR	Double false discovery rate[1]
GBH	Yes	FDR	Group Benjamini–Hochberg[2]
BH	No	FDR	False discovery rate control by the BH-procedure[18]
NOADJ	No	Individual hypothesis	Unadjusted hypothesis testing

Comparison studies have been performed before on some of these methods.[1,5,19] A number of these studies compare power and error rates for detecting increases in AE rate on the treatment arm for Bayesian methods against two-sided hypothesis (Fisher) tests for the error controlling procedures,[5] effectively ignoring the possibility of a decrease in AE rate on the treatment arm for the Bayesian methods, and any associated errors. The individual AE rates are generally assumed to be sampled from constant underlying SOC AE rates. Typically there is more variation in real AE incidence rates within SOCs than is often considered in these comparison studies. For the error controlling procedures operating on *p*-values this is not an issue, but for the Bayesian methods, where the AE incidence model is centred around SOC mean values, we investigate if the models are flexible enough to cope with this variation. Our approach is designed to address some of the gaps in previous studies. It systematically compares both the error controlling methods and Bayesian models while taking directly into account the variability in AE occurrence rate within SOCs, the impact of changes in the data structure by removing low incidence AEs or complete SOCs, and the inclusion of decreases in AE rate on the treatment arm, allowing a general balanced comparison of the methods. While the analyses are performed with the purpose of investigating some of the statistical properties of the methods, clinical opinion remains of primary importance in determining any safety issues.

2 | METHODS

The methods in this study (Table 1) may be divided into two categories: direct error control procedures and Bayesian data modelling. The error controlling approaches included are control of the FDR by the Benjamini–Hochberg procedure (BH),[18] the double false discovery rate (DFDR),[1] and the Group Benjamini–Hochberg procedure (GBH)[2]; and unadjusted hypothesis testing (NOADJ). The modelling approaches are the hierarchical Bayesian model of Berry and Berry (BB),[3] where the prior distribution of an increase or decrease in log-odds of an AE occurring on the treatment arm (θ) is modelled by a mixture distribution consisting of a point-mass at zero and a normal distribution, and a similar model without the point-mass term (1a).[5] A θ value of 0 indicates no difference between the trial arms. This corresponds to the null hypothesis for a standard test (e.g., Fisher exact test). A large posterior probability of an increase or decrease in θ is an indication of a difference between trial arms and is the method of assessing differences in treatment rate for the Bayesian methods in this study.

For the error controlling procedures we use a standard 5% significance level for unadjusted hypothesis testing and 5% and 10% (error) levels for methods which control the FDR. The 10% level is included as this is the recommended level for the DFDR.[1] The hypothesis test used is the Fisher exact test which is typical for end of trial AE incidence data.

The direct comparison of error controlling procedures and Bayesian models raises a number of points. While comparisons between the error controlling procedures at particular (error) levels are possible, there is no standard method of comparing with the Bayesian approaches. As we are interested in assessing the utility of the methods for determining which AEs are associated with treatment, comparisons between the numbers of AEs correctly detected, and the associated error rates, are needed. Although power and FDR may be considered classical concepts, based on the posterior probability of an increase or decrease in the log odds on the treatment arm, we can estimate their equivalent values for the Bayesian methods, provided the underlying AE generating process is known. This requires that a suitable threshold value be chosen for the Bayesian methods above which an AE is declared associated with treatment. A further issue is that it not immediately obvious how to choose a threshold which corresponds to a specific target FDR rate. Theoretical approaches to determining thresholds have been suggested,[20] but, based on previous studies,[16,17,19] we use as the event flagging mechanism for the Bayesian models threshold values of 0.975 and 0.95 for model 1a, and 0.90 and 0.80 for model BB. We chose two values for each model in order to give some idea of the how the estimated power and FDR vary as the threshold values change, and to allow a more accurate assessment when comparing to the error controlling procedures. Model BB threshold values are lower than those for model 1a as model BB requires a stronger signal to overcome the barrier provided by the point-mass term. A further important point is that the error controlling procedures generally use a test which is invariant under switching of treatment and control data. However, this may not be the case for modelling approaches where the structure of the model and choice of data set as control or treatment may have an effect on the results.

3 | CASE STUDY

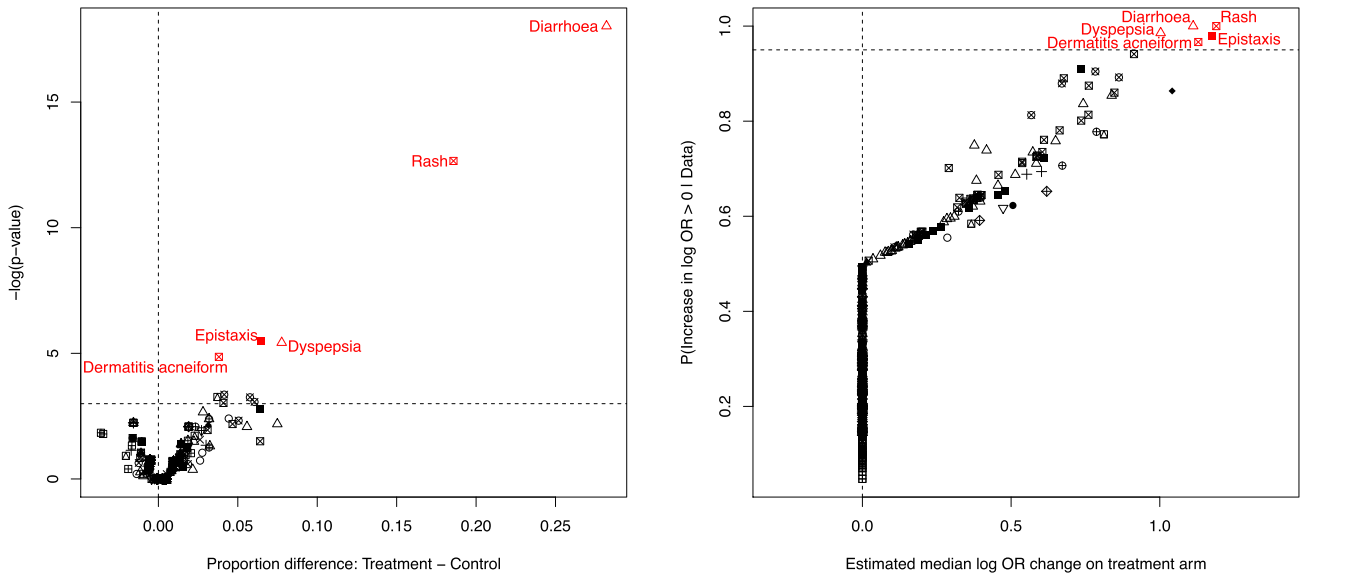
The case study is based on a GlaxoSmithKline plc (GSK) sponsored Phase III randomised clinical trial.† The AE data, trial descriptions, clinical study reports, and a number of related publications, which reflect various aspects of the trial,

are available through the GSK clinical study register.‡ *Diarrhoea* was designated an AE of special interest in the trial protocol, and a number of other AEs, such as *Rash*, were expected to have raised treatment rates. There were 191 patients on the control arm and 210 patients on the treatment arm and incidence counts for 497 different AEs, spread over 23 SOCs, were recorded during the trial.

Two issues are immediately apparent. Firstly, the large numbers of SOCs and AEs are likely to make it difficult for the error controlling procedures to flag any but the AEs with the largest differences between treatment and control. This is not a problem for the Bayesian methods where it can be argued that the inclusion of all the recorded AEs is an important part of the modelling process.[3] International Conference on Harmonisation (ICH) guidelines[13] allow for the possibility of removal of low severity or low count AEs and the DFDR requires an initial dimension-reduction step 'that disregards AE types for which the total incidence is so low (rare AEs) that statistical significance at the conventional 0.05 level is impossible even without a multiplicity adjustment'.[1] Xia el al. detail this step as being the removal of AEs with counts of less than 5 in both groups combined.[5] In this trial 172 of the AEs recorded affected more than 1% of the patients on at least one trial arm, and 120 AEs had a total incidence of greater than 5 on both trial arms combined. While clinical expertise about AE inclusion in the statistical analysis is of primary importance, scanning or exploration of all the safety data has the potential to highlight unexpected AEs which could be potential future issues.[21,22] Of the methods in this study, only the DFDR has a requirement to remove low incidence AEs before the analysis. Secondly, the SOC *General disorders and administration site conditions* is designated by MedDRA to be used for AEs which are nonspecific or that may be related to several SOCs. If the AEs in this SOC have no clinical relationship with each other then an analysis that uses information sharing may be counterproductive. This is particularly so for the Bayesian methods. For the error controlling procedures this may reduce the power but control of the FDR should remain at the required level.

Initially a two-sided Fisher exact test is applied to compare the AEs on different trial arms (NOADJ). A plot of the treatment differences versus $-\log(p\text{-value})$ is shown in Figure 1(A) indicating that in addition to *Diarrhoea* and *Rash* a number of other of AEs may have increased incidence on the treatment arm. Ten AEs in total were significant at the 5% level for the test and overall the treatment was considered to be well tolerated. The AEs with the lowest *p*-values came from a small number of SOCs, indicating a possible SOC effect. The remaining methods in Table 1 were applied and the results for the 10 significant AEs from the Fisher test are given in Table 2.

The application of the error controlling procedures to the full set of trial safety data leaves only *Diarrhoea* and *Rash* as significant at both the 5% and 10% levels. Even *Epistaxis* and *Dyspepsia*, which have what could be considered small



(A) Treatment differences versus -log(p-value).

(B) BB: Plot of median change in log odds-ratio (OR) under treatment versus the posterior probability of an increase.

FIGURE 1 Treatment differences. Each SOC is represented by a different symbol. (A)Treatment differences versus $-\log$ (*p*-value). (B) BB: Plot of median change in log odds-ratio (OR) under treatment versus the posterior probability of an increase

p-values, are not flagged. The inclusion of large numbers of very low incidence AEs make it impossible for the error controlling procedures to flag any but the strongest signalled AEs when controlling for multiple hypotheses, whether using standard or grouped approaches. In contrast, the Bayesian methods are not as sensitive to the inclusion of very low incidence AEs.[3] Using a cut-off point of 0.90 posterior probability we would have flagged seven events using the Berry and Berry model (BB). For a cut-off of 0.95 this drops to five events. Figure 1(B) shows the gradual fall-off in posterior probabilities (vs. the change in median log odds ratio) for the fitted BB model.

There is also clear evidence of the effect of the SOC on the results for the BB model. Of the 10 AEs in Table 2 only 9 are in the top 10 by posterior probability for the BB model. *Localised infection*, in the *Infections and infestations* SOC has a posterior probability of an increase in odds-ratio of 0.772, and has been replaced by *Dyspnoea* from the *Respiratory, thoracic and mediastinal disorders* SOC (posterior probability: 0.896). For *Localised infection* as there are no other AEs in that SOC with high occurrence rates on the treatment arm, the model has shrunk the effect. For *Dyspnoea* on the other hand, *Epistaxis* is also in the *Respiratory, thoracic and mediastinal disorders* SOC, and here there is evidence of borrowing strength. For model 1a the situation is somewhat different. All 10 events in Table 2 have high posterior probabilities and overall 19 events have posterior probability greater than 0.975. The posterior probabilities are higher compared to the BB model. This is due to the presence of the point-mass term in BB which requires a stronger signal in order to detect differences between treatment and control.

The situation changes as low incidence AEs are removed from the data. For example the posterior probability for *Localised infection* for the BB model drops to 0.634 when AEs affecting less than 1% of patients on both trial arms are removed, indicating it is unlikely to be associated with treatment. With the removal of low incidence AEs there has been a corresponding increase in the number of AEs flagged by the DFDR and GBH, although some of these are only at the 10% level (Table 2). Overall the methods give results similar to the trial conclusions but with the addition of a number of AEs for consideration highlighted by some of the methods. How important these additional AEs may be is of course dependent on considerations beyond statistical modelling.

The Bayesian approaches for scanning the data do not appear to be as sensitive to the changing size of the data as the error controlling procedures. However, removing AEs from the SOC hierarchy does have an effect on the results with the posterior probabilities changing, reflecting the altered compositions of the SOCs. For the error controlling procedures, the grouped methods only start becoming more effective than their non-grouped equivalents when low-incidence AEs are removed. We investigate further the effect of removing low count AEs on all the methods in the following comparative study.

TABLE 2 Trial EGF100151: Adverse events significant at 5% for Fisher exact tests

SOC	Adverse event	NOADJ	BB[a]	1a[a]	BH[b]	DFDR[b]	GBH[b]
Gastrointestinal disorders	Diarrhoea	<0.001	1.000	1.000	Y	Y	Y
Skin and subcutaneous tissue disorders	Rash	<0.001	1.000	1.000	Y	Y	Y
Respiratory, thoracic and mediastinal disorders	Epistaxis	0.004	0.980	0.999	N	N	Y[c]
Gastrointestinal disorders	Dyspepsia	0.004	0.986	0.999	N	Y[d]	Y[d]
Skin and subcutaneous tissue disorders	Dermatitis acneiform	0.008	0.967	0.995	N	Y[e]	Y[e]
Musculoskeletal and connective tissue disorders	Muscle spasms	0.035	0.892	0.987	N	N	N
Infections and infestations	Localised infection	0.038	0.772	0.983	N	N	N
Musculoskeletal and connective tissue disorders	Arthralgia	0.039	0.905	0.993	N	N	N
Musculoskeletal and connective tissue disorders	Back pain	0.047	0.879	0.991	N	N	N
Skin and subcutaneous tissue disorders	Nail disorder	0.049	0.941	0.993	N	N	N

[a]Posterior probability that the change in log-odds of the occurrence of the adverse event on the treatment arm is positive.

[b]Flagged as significant at the 5% level unless otherwise indicated (Y = yes, N = no).

[c]Flagged as significant at the 5% level for AEs affecting more than 1% on both arms and for AEs with combined arms total greater than 5.

[d]Flagged as significant at the 10% level for AEs affecting more than 1% on both arms. Flagged as significant at the 5% level for AEs with combined arms total greater than 5.

[e]Flagged as significant at the 10% level for AEs affecting more than 1% on both arms. Flagged as significant at the 10% level for AEs with combined arms total greater than 5.

4 | COMPARATIVE STUDY

The comparative study is based on the case study trial (497 AEs/23 SOCs). The control probabilities are estimated from the original trial data, with zero occurrence AEs on the control arm given an incidence probability of 0.002. We present the results from two scenarios. In the first (CS1) 10 AEs from 5 SOCs have changed treatment arm rates. *Diarrhoea* and *Rash* have an absolute increase of 0.15 in the probability of an event on the treatment arm. *Dyspepsia*, *Nail disorder*, *Dermatitis acneiform*, *Localised infection* and *Epistaxis* have an increase of 0.05. *Back pain*, *Arthralgia* and *Muscle spasms* have a decrease of 0.05 on the treatment arm. All other AEs have equal probability of occurrence on both control and treatment arms. In the second (CS2), in addition to the 10 AEs with changed rates in CS1, approximately one quarter of AEs in the 5 affected SOCs have changed rates, giving 36 AEs in total with different rates between the trial arms. Full details of the simulation study and additional results are given in the Supporting Information.

Comparisons between the methods focuses on the numbers of AEs correctly identified as having different treatment rate, and on the misclassification of AEs as measured by the FDR. Given the effect that including large numbers of low incidence AEs has on the methods in the case study, four analyses are made on each data set: the first on the full set of data (*All AEs*), the second on a reduced set of data where AEs whose occurrence rates were less than 1% on each trial arm were excluded (*AEs* [*>1%*]), the third where only AEs with combined totals from the trial arms of more than 5 AEs are included (*AEs* [*>5*]), and the fourth where 10 SOCs are removed from the analysis (*Removed SOCs*). The results for all methods for the full and reduced data sets are given in Tables 3 and 4.

For both CS1 and CS2 removing SOCs which do not contain AEs with changed treatment rates (*Removed SOCs*) does not generally affect the results for the grouped methods, with power and FDR similar to the full data set. In both simulations the power of the Bayesian methods remains relatively constant (for each threshold) despite the changes in data size, but, in particular in CS1 (*AEs* [*>1%*]), the FDR is clearly affected by the removal of the low incidence AEs. Removing low incidence AEs removes the shrinkage effect of these AEs within their SOCs. The power of the error controlling procedures increases as low incidence AEs are removed. The GBH is the most powerful of these methods but as it only guarantees asymptotic error control, the estimated FDR increases above nominal levels as low incidence AEs are removed.

For CS1 full data set (*All AEs*), model 1a at the 0.95 level correctly detects the most AEs (66.26%) but with a correspondingly high FDR (34.17%), dropping to 54.56% and 18.75% when the threshold is moved to 0.975. Model BB at the 0.80 level has the third highest power (41.19%) but a much tighter control on the FDR (10.55%), dropping to 31.25% and 4.05%, respectively, at the 0.90 threshold. In comparison the error controlling approaches have not performed as well with GBH at the 0.10 level performing best (power: 31.05%, FDR: 6.73%). The DFDR performs poorly at both the 0.05 and 0.10 levels, and does not outperform the BH procedure in terms of power or control of the FDR. The results are

TABLE 3 CS1: Estimated FDR and power

Method	Level	All AEs		AEs (>1%)		AEs (>5)		Removed SOCs	
		FDR (%)	Power (%)	FDR (%)	Power (%)	FDR (%)	Power (%)	FDR (%)	Power (%)
1a	0.950	34.17	66.26	46.10	63.59	34.99	63.28	26.17	66.54
1a	0.975	18.75	54.56	22.23	51.52	19.78	50.87	13.75	54.70
BB	0.80	10.55	41.19	24.21	40.50	9.18	38.38	9.46	43.72
BB	0.90	4.05	31.25	7.04	29.97	3.19	28.52	3.46	33.23
GBH	0.05	4.13	24.62	8.13	32.99	12.27	37.41	2.34	24.62
GBH	0.10	6.73	31.05	15.04	40.22	22.76	45.23	4.05	31.05
DFDR	0.05	1.34	11.74	2.63	18.12	4.22	22.48	1.04	14.34
DFDR	0.10	2.30	16.56	5.14	24.34	9.06	30.30	1.75	19.64
BH	0.05	0.57	12.17	1.42	18.77	2.61	23.13	0.40	14.76
BH	0.10	0.96	16.47	3.14	25.61	5.26	30.81	1.10	19.83
NOADJ	0.05	36.42	62.58	36.42	62.58	36.42	62.58	27.92	62.58

Abbreviations: 1a, Berry and Berry model without point mass; BB, Berry and Berry model; BH, Benjamini–Hochberg procedure; DFDR, double false discovery rate; GBH, Group Benjamini–Hochberg procedure; NOADJ, unadjusted hypothesis testing.

TABLE 4 CS2: Estimated FDR and power

Method	Level	All AEs		AEs (>1%)		AEs (>5)		Removed SOCs	
		FDR (%)	Power (%)	FDR (%)	Power (%)	FDR (%)	Power (%)	FDR (%)	Power (%)
1a	0.950	13.78	83.33	15.73	82.69	13.13	84.11	11.34	83.58
1a	0.975	6.27	75.24	6.18	73.38	6.33	75.69	5.08	75.65
BB	0.80	6.55	73.02	9.18	76.22	5.73	75.04	6.54	73.96
BB	0.90	2.36	63.40	2.78	65.32	2.33	64.31	2.25	64.32
GBH	0.05	0.95	39.21	2.45	50.52	4.16	57.28	0.66	39.21
GBH	0.10	2.10	49.39	5.67	61.08	8.89	68.31	1.58	49.39
DFDR	0.05	0.30	28.48	1.07	41.82	1.81	49.36	0.40	32.78
DFDR	0.10	1.03	40.67	2.74	54.94	4.58	62.09	0.96	44.63
BH	0.05	0.38	23.89	0.97	36.73	1.74	44.16	0.41	30.02
BH	0.10	0.78	34.20	2.32	48.84	4.16	56.21	0.82	41.27
NOADJ	0.05	10.52	70.74	10.52	70.74	10.52	70.74	7.02	70.74

Abbreviations: 1a, Berry and Berry model without point mass; BB, Berry and Berry model; BH, Benjamini–Hochberg procedure; DFDR, double false discovery rate; GBH, Group Benjamini–Hochberg procedure; NOADJ, unadjusted hypothesis testing.

similar for the reduced data set *AEs* (*>1%*) but here the FDR for the Bayesian method is not well controlled with the FDR for model 1a at the 0.95 level being 46.19%. Similarly the FDR for BB at level 0.80 is 24.21%, comparing poorly with the error controlling methods, although for the GBH at level 0.10 the FDR is 15.05%. The power of the error controlling methods increases in all cases compared to the full data set. For the data set *AEs > 5* the FDRs of the Bayesian methods has reverted to levels similar to the full data set. For CS2 the results are broadly similar to CS1 but with increased power as might be expected, with a stronger SOC effect due to the increased numbers of AEs with changed rates in their relevant SOCs. Here in all cases there is better control of the FDR and the DFDR shows increased performance compared to BH in this data set.

Assessing the methods requires that a balance be struck between power and error control. For the full data sets (*All AEs*) and data sets with removed SOCs (*Removed SOCs*), provided a suitable threshold has been chosen, the Bayesian modelling approaches appear to have performed best overall in terms of power and control of the FDR, with model 1a even out performing unadjusted testing (NOADJ). In this study the power of the Bayesian methods is more stable over the different data sizes than those of the error controlling procedures. However, due to the shrinkage effects of the models, control of the FDR may be affected by how low incidence AEs are removed. The GBH appears to be best performing of the error controlling methods but does not maintain tight control of the FDR as the data size decreases. This is not unexpected, the GBH only claims control of the FDR asymptotically.[2] The DFDR does not perform well for CS1, where it is no more powerful than the standard BH procedure, but outperforms it in CS2.

5 | DISCUSSION

Clinical input is of primary importance in recognising which AEs are associated with treatments, but routinely applying statistical methods to the large quantities of AE data available in a clinical trial allows the possibility of flagging unexpected AEs for investigation at the pre-marketing phase of treatment development. The comparative study adds evidence to the belief that where the underlying assumptions of grouped AEs are true the use of grouped methods is appropriate, and also adds insight as to which methods are most affected by removal of low count AEs, or SOCs not considered to be of clinical interest.

While most AEs recorded during a clinical trial will not have a specific associated hypothesis, for information sharing (Bayesian) methods, their inclusion may have an important role in the analysis. We see this for *Localised infection* in the case study, where the posterior probability under the BB model drops as AEs are systematically excluded from the analysis. We have also seen in the comparative study that removing AEs based on rules can have an impact on the FDR for the Bayesian models. On the other hand the inclusion of many low count AEs reduced the effectiveness of

the error controlling procedures, and removing them improved the performance in terms of power, again however it can lead to increases in the FDR. For this reason the systematic removal of AEs from the data included for analysis based on low incidence rates alone is not recommended. However, the removal of SOCs not deemed to be of clinical interest does not have a dramatic effect on the results of any of the methods.

One advantage of the Bayesian approaches is that they perform well even when including the large numbers of AEs recorded during the trial and in fact may be adversely affected in terms of error rate by the removal of low incidence AEs as opposed to the removal of complete SOCs. So while overall, provided care is taken, the Bayesian methods may provide a better power/error balance than the corresponding error controlling procedures, the structure of the data included in the analysis plays an important role. In particular the structure of the individual SOCs in terms of the AEs included in the analysis, rather than the inclusion of the SOCs themselves, is a key part of AE analysis.

While the results presented and discussed here are empirical, we can consider if and how they may be justified theoretically, based on an understanding of the method definitions. Looking first at the removal of low incidence AEs, for the Bayesian methods, where borrowing strength is a key part of the model approaches, there is some possibility of an increase in FDR. The presence of low incidence AEs is expected to shrink the model effects and removing them may result in some AEs being incorrectly flagged as associated with treatment. The potential impact of the removal of low incidence events on power could result in an increase, due to the loss of the shrinkage effect, but it is also possible that the removed AEs would have been flagged as associated with treatment, even with low numbers. In this case the power would decrease. In our simulations the effect on power was minimal, but while the FDR is similar for *AEs* (*>5*) as for *All AEs*, the FDR is very much larger for *AEs* (*>1%*). An important point here is that for *AEs* (*>1%*) the number of AEs excluded and their values are dependent on the size of the trial, so quantifying this is difficult. For the GBH we expect an increase in FDR as AEs are removed, control of the FDR by the GBH at the specified level is only guaranteed asymptotically. Power and FDR may also be expected to increase as the estimation of the proportion of significant AEs in each SOC should increase with the low incidence AEs removed.[2] For the DFDR, where removal of low incidence AEs is a requirement of the method, we also expect to see an increase in power. The DFDR is a two-step procedure with a mechanism for including or excluding a SOC in a final selection of candidate AEs. For simulation CS1, the low number of AEs associated with treatment and the large number of SOCs mitigate against AE detection. For CS2, with the increased numbers of AEs associated with treatment, the method performs much better.

The removal of SOCs which do not contain AEs associated with treatment (*Removed SOCs*) would not be expected to have a large impact on the power of the Bayesian model results. The SOC relationships in the model are weak. However, their removal has the potential to reduce the FDR as these AEs can no longer be flagged as incorrectly associated with treatment. For the GBH, SOCs with no AEs associated with treatment would be unlikely to contain any flagged AEs, so we would expect very similar power and FDR rates compared to the full data set. For the DFDR, where the method is more dependent on the number of SOCs in the analysis, we would expect and see increased power.

6 | CONCLUSION

Compared to the post-marketing analysis of AE incidence data, a safety analysis during a clinical trial may take advantage of the balance between comparator groups to make causal inferences about the association of AEs with treatments. Unlike post-marketing analyses, which may have access to data on large numbers of patients over long time periods, clinical trials are of limited duration and hence are necessarily limited in detecting rare or possibly unexpected events. Grouped methods are one approach to addressing this limitation.

This study has looked at the application of a number of grouped methods to clinical trial safety data, primarily from a statistical point of view. However, the study has a number of limitations which must be acknowledged. The results are empirical, and necessarily restricted to the data sets analysed. The determination of a suitable threshold for the Bayesian analysis was not part of the study, and relating this to the corresponding FDR level for the error controlling procedures was not straightforward. Two thresholds for each Bayesian model were chosen to give a better view of how the methods compared. A major part of the grouped method approach is the assumption that there are relationships between the AEs and that these can be captured at a SOC or similar grouping level. The nature of these relationships may be more complicated than that considered in this study.

However the study has expanded the comparisons of the methods beyond those covered in the original papers, with more realistic trial sizes and more variable data, covering at least some of the analyses which may arise in reality. The results presented are also in line with what might be expected theoretically, allowing a number of general conclusions

to be drawn, while still acknowledging the limitations of what can be achieved in a study such as this. In particular, the structure of the groupings of AEs by SOC plays an important role both in the detection of AEs associated with treatment, and the error rates of the various methods, with the inclusion or removal of low incidence AEs in the analysis having an effect on the results. The determination of which AEs to include in the trial safety analysis may not be a straightforward task, but should be defined before the analysis. While clinical input is of primary importance, the desire to detect rare events in the clinical trial phase requires careful consideration of the AE data included in the safety analysis, even if some AEs are not considered to be clinically relevant. This issue tends not to occur in the post-marketing phase where analysis is often restricted to those AEs which have occurred as opposed to those which might occur.

Safety analysis in clinical trials is a complex problem which continues to be an area of interest to researchers with new methods continuing to be developed.[23–25] The methods compared here are primarily suited to an end of trail incidence analysis, and the availability of software implementations[11,12] has the potential to allow the integration of the methods into a trial's safety analysis with relatively small overhead, and is something that should be considered for inclusion.[26]

ACKNOWLEDGMENTS

This work was supported by the Engineering and Physical Sciences Research Council (UK) (EPSRC) [award reference 1521741] and Frontier Science (Scotland) Ltd., and by Health Data Research (UK) [award reference MR/S003967/1].

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

ENDNOTES

* http://www.meddra.org/.

† ClinicalTrials.gov identifier: NCT00078572.

‡ http://www.gsk-clinicalstudyregister.com, ID: EGF100151.

DATA AVAILABILITY STATEMENT

The supplementary material for this paper contains details of the software and simulations used in the paper as well as a number of additional scenarios. The scripts used to run the analyses are available from the corresponding author.

ORCID

Raymond Carragher https://orcid.org/0000-0002-0120-625X

REFERENCES

1. Mehrotra DV, Adewale AJ. Flagging clinical adverse experiences: reducing false discoveries without materially compromising power for detecting true signals. *Stat Med*. 2012;31(18):1918-1930. https://doi.org/10.1002/sim.5310
2. Hu JX, Zhao H, Zhou HH. False discovery rate control with groups. *J Am Stat Assoc*. 2010;105(491):1215-1227. https://doi.org/10.1198/jasa.2010.tm09329.
3. Berry SM, Berry DA. Accounting for multiplicities in assessing drug safety: a three-level hierarchical mixture model. *Biometrics*. 2004;60(2):418-426. https://doi.org/10.1111/j.0006-341x.2004.00186.x.
4. DuMouchel W. Multivariate Bayesian logistic regression for analysis of clinical study safety issues. *Stat Sci*. 2012;27(3):319-339. https://doi.org/10.1214/11-sts381.
5. Amy Xia H, Ma H, Carlin BP. Bayesian hierarchical modeling for detecting safety signals in clinical trials. *J Biopharm Stat*. 2011;21(5):1006-1029. https://doi.org/10.1080/10543406.2010.520181.
6. Crooks CJ, Prieto-Merino D, Evans SJW. Identifying adverse events of vaccines using a Bayesian method of medically guided information sharing. *Drug Saf*. 2012;35(1):61-78. https://doi.org/10.2165/11596630-000000000-00000.
7. Wang W, Whalen E, Munsaka M, et al. On quantitative methods for clinical safety monitoring in drug development. *Stat Biopharm Res*. 2018;10(2):85-97. https://doi.org/10.1080/19466315.2017.1409134.
8. Munsaka MS. A question-based approach to the analysis of safety data. In: Peace Karl E, Ding-Geng C, Sandeep M, eds. *Biopharmaceutical Applied Statistics Symposium: Volume 2 Biostatistical Analysis of Clinical Trials*. Singapore: Springer Singapore; 2018:193-216.
9. Melvin M. *A Question-Based Approach to the Analysis of Safety Data: Volume 2 Biostatistical Analysis of Clinical Trials*. Singapore: Springer; 2018:193-216.
10. Fries M, Kracht K, Li J. Safety monitoring methodology in the premarketing setting. *Proceedings of JSM*. 2016:2247-2269.

11. Carragher R. c212: Methods for Detecting Safety Signals in Clinical Trials Using Body-Systems (System Organ Classes); 2017. https://CRAN.R-project.org/package=c212
12. Carragher R, Robertson C. c212: An R package for the detection of safety signals in clinical trials using body-systems (system organ classes). *J Open Source Softw*. 2020;5(56):2706. https://doi.org/10.21105/joss.02706.
13. International Conference on Harmonisation E9 Expert Working Group. Statistical principles for clinical trials. ICH harmonised tripartite guideline. *Stat Med*. 1999;18(15):1905-1942.
14. Gelman A, Hill J, Yajima M. Why we (Usually) don't have to worry about multiple comparisons. *J Res Educ Effect*. 2012;5(2):189-211. https://doi.org/10.1080/19345747.2011.618213.
15. Dunson DB, Herring AH, Engel SM. Bayesian selection and clustering of polymorphisms in functionally related genes. *J Am Stat Assoc*. 2008;103(482):534-546. https://doi.org/10.1198/016214507000000554.
16. Odani M, Fukimbara S, Sato T. A Bayesian meta-analytic approach for safety signal detection in randomized clinical trials. *Clin Trials*. 2017;14(2):192-200. https://doi.org/10.1177/1740774516683920.
17. Carragher R, Mueller T, Bennie M, Robertson C. A Bayesian hierarchical approach for multiple outcomes in routinely collected healthcare data. *Stat Med*. 2020;39(20):2639-2654. https://doi.org/10.1002/sim.8563.
18. Yoav B, Yosef H. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc Ser B*. 1995;57(1):289-300.
19. Carragher R. Detecting Safety Signals in Randomised Controlled Trials. PhD thesis. University of Strathclyde; 2017.
20. Chen W, Zhao N, Qin G, Chen J. A Bayesian group sequential approach to safety signal detection. *J Biopharm Stat*. 2013;23(1):213-230. https://doi.org/10.1080/10543406.2013.736813.
21. Southworth H. Predicting potential liver toxicity from phase 2 data: a case study with ximelagatran. *Stat Med*. 2014;33(17):2914-2923. https://doi.org/10.1002/sim.6142.
22. Siddiqui O. Statistical methods to analyze adverse events data of randomized clinical trials. *J Biopharm Stat*. 2009;19(5):889-899. https://doi.org/10.1080/10543400903105463
23. Tan X, Liu GF, Zeng D, et al. Controlling false discovery proportion in identification of drug-related adverse events from multiple system organ classes. *Stat Med*. 2019;38(22):4378-4389. https://doi.org/10.1002/sim.8304.
24. Tan X, Chen BE, Sun J, Patel T, Ibrahim JG. A hierarchical testing approach for detecting safety signals in clinical trials. *Stat Med*. 2020;39(10):1541-1557. https://doi.org/10.1002/sim.8495.
25. Diao G, Liu GF, Zeng D, et al. Efficient methods for signal detection from correlated adverse events in clinical trials. *Biometrics*. 2019;75(3):1000-1008. https://doi.org/10.1111/biom.13031.
26. Carragher R. Comparative study for: "Assessing safety at the end of clinical trials using system organ classes: a case and comparative study"; 2020.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Carragher R, Robertson C. Assessing safety at the end of clinical trials using system organ classes: A case and comparative study. *Pharmaceutical Statistics*. 2021;1–10. https://doi.org/10.1002/pst.2148