Received: 21 June 2019 | Revised: 10 April 2020 | Accepted: 13 April 2020

DOI: 10.1002/sim.8563

RESEARCH ARTICLE

Statistics in Medicine WILEY

A Bayesian hierarchical approach for multiple outcomes in routinely collected healthcare data

Raymond Carragher[1,2,3] | **Tanja Mueller**[1] | **Marion Bennie**[1,4] | **Chris Robertson**[2,5]

[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 and Intelligence Strategic Business Unit, NHS National Services Scotland, Glasgow, UK

[5]Health Protection Scotland, NHS National Services Scotland, Glasgow, UK

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

Clinical trials are the standard approach for evaluating new treatments, but may lack the power to assess rare outcomes. Trial results are also necessarily restricted to the population considered in the study. The availability of routinely collected healthcare data provides a source of information on the performance of treatments beyond that offered by clinical trials, but the analysis of this type of data presents a number of challenges. Hierarchical methods, which take advantage of known relationships between clinical outcomes, while accounting for bias, may be a suitable statistical approach for the analysis of this data. A study of direct oral anticoagulants in Scotland is discussed and used to motivate a modeling approach. A Bayesian hierarchical model, which allows a stratification of the population into clusters with similar characteristics, is proposed and applied to the direct oral anticoagulant study data. A simulation study is used to assess its performance in terms of outcome detection and error rates.

KEYWORDS

Bayesian hierarchy, direct oral anticoagulants, multiple outcomes, observational study, safety outcomes

1 | INTRODUCTION

Clinical trials remain the standard method for establishing the efficacy and safety of new treatments.[1] While clinical trials are able to provide a causal analysis of the relationship between treatments and outcomes, they also have a number of potential limitations. If the trial population does not reflect the general population on which the treatment is actually used once approved there may be some concern about the generalizability of the trial results;[2] a trial may only compare a small number of possible treatments for any particular condition, even if many such treatments are available; furthermore, as the trial is generally sized to answer a primary objective, the power to detect differences for any additional hypotheses, including the detection of safety issues, may be reduced. For approved treatments, postmarketing surveillance provides longer term safety analysis outside of the trial environment. A number of regulatory agencies and drug monitoring centers have developed computerized methods for identifying potential serious adverse outcomes using spontaneous report adverse event databases.[3] For example, the US Food and Drug Administration (FDA) use a Bayesian data-mining approach[4] and the World Health Organization (WHO) use a Bayesian neural network.[5]

Assessing the generalizability of treatments to populations not covered by clinical trials, or indeed assessing new treatments in comparison with preexisting treatments, may be difficult, if not impossible, based on clinical trial

results alone, and will require the use of nontrial data beyond that provided by adverse event reporting databases. These types of analyses may be considered part of a comparative effectiveness research (CER) approach which seeks to compare healthcare interventions to determine which work best in the sense that they provide the most benefits and the least harm to patients, bringing the prospect of a precision medicine approach for individual patients closer.

Sources of information beyond trial data and adverse event reporting databases are becoming increasingly available to healthcare researchers. In particular many national healthcare providers record information about patient demographics, issued prescriptions, treatment duration, adherence, co-morbidities, hospitalizations, and other outcomes of different treatment regimes. These records together with national registers of births and deaths provide a source of information regarding the relative performance of different treatments in the general population. This data are constantly accumulating as new patients interact with the relevant health bodies.

The analysis of this type of observational data provides a number of logistical and statistical challenges. Data may come from multiple sources with conflicting or contradicting information and may require significant clean-up before it can be analyzed. Patient outcomes, either recorded as ICD-10[1] codes or locally coded, may not map directly to adverse event definitions as defined by clinical trials, making comparisons with trial outcomes difficult. So while we may expect to see similar patterns of outcomes for patients under treatment in the general population as we have seen in clinical trials, the lack of a direct general mapping between the two sets of outcomes will make this more difficult to determine.

The structure and balance provided by clinical trials does not exist in observational data and the analysis of the data may be inhibited by bias, for example, by treatment indicator. The use of methods such as propensity score analysis or inverse probability of treatment weighting (IPTW), under certain assumptions, is possible approaches to handling bias, but there is no objective method for defining a causal analysis in an observational study.[6] More recent developments using genetic indicators or similar instrumental variables have the potential to provide a "quasi-randomization" of the observed populations, opening the possibility of more balanced treatment comparisons.[7]

In observational studies, outcomes are often modeled individually. This has a number of statistical implications. Modeling individually excludes the possibility of using relationships, which may exist between the outcomes in the analysis, and there is no straightforward way of assessing the impacts of the outcomes together in order to make a decision about which treatments are suitable for particular patients. There is also the potential lack of control for multiple comparisons, leading to the possibility of spurious associations being detected in the data.

In clinical trials patient outcomes, such as adverse events, may be defined by specific clinical symptoms or laboratory measurements, and may be the clinical expression of the effect a drug has on a particular organ or body-system. For example, in the United States National Cancer Institute (NCI) Common Terminology Criteria for Adverse Events (CTCAE),[2] adverse events such as *cholecystitis* and *hepatic pain* are defined as part of the *hepatobiliar disorders* grouping of events and would typically be reported separately. Most medical dictionaries, for example, MedDRA or WHO-ART, define a hierarchical structure consisting of system organ classes (SOCs), various groupings (higher level terms), and descriptor or preferred terms for describing the adverse event itself. In clinical trial, study reports safety data are often presented grouped by system organ class and methods, which take advantage of this type of grouping approach have been developed for adverse event analysis.[8-10] These methods assume that in a SOC affected by a treatment we may be more likely to see raised occurrence rates for a number of related adverse events within that SOC and the analysis tries to take advantage of this extra information when modeling the data. While ICD-10 codes do provide a hierarchy of sorts for nontrial patient data, the groupings of outcomes is not as well defined as those supplied in medical dictionaries and generally used in trials.

In Section 2, an existing study of direct oral anticoagulants is introduced and a number of the characteristics that affect the analysis of this type of data are discussed and used to motivate a Bayesian hierarchical modeling approach based on the adaption of existing clinical trial methods. In Section 3, we extend the some of these clinical trial approaches to allow multiple treatments and outcomes, with patients grouped into stratified clusters, for use in an observational setting. The methods are applied to a simulated study to assess their performance and error rates (Section 4.1) and then to data from the study of direct oral anticoagulants in Scotland (Section 4.2).

[1]International Classification of Disease codes, 10th edition.
[2]CTCAE Version 4.0: http://evs.nci.gov/

TABLE 1 DOAC prescription characteristics for the Forth Valley health board

Health board	Treatment	First recorded prescription (days since drug approval)[a]	Last recorded prescription (days to end of study)[b]
Forth Valley	Dabigatran	10/11/2011 (97)	31/08/2014 (497)
Forth Valley	Apixaban	07/08/2013 (208)	18/12/2015 (13)
Forth Valley	Rivaroxaban	02/05/2012 (110)	30/12/2015 (1)

[a]For example, the 10/11/2011 was the date of the first recorded Dabigatran prescription in the Forth Valley health board, and this was 97 days after the drug was approved for use.
[b]For example, the 31/08/2014 was the date of the last recorded Dabigatran prescription in the Forth Valley health board, and this was 487 days before the end of the study date.

2 | CASE STUDY: DIRECT ORAL ANTICOAGULANT SCOTLAND STUDY (2011-2015)

The DOAC Scotland Study (2011-2015) is a study of the comparative safety and effectiveness of direct oral anticoagulants (DOACs) in Scotland for patients with a hospital confirmed diagnosis of atrial fibrillation, from August 2011 to December 2015.[2] The study population consisted of 14788 patients initiating one of three treatments: Apixaban, Dabigatran, or Rivaroxaban. The data were analyzed using a number of different approaches, including Cox regression on the index treatment (the first treatment received), with censoring on patient death, treatment discontinuation, or treatment switch. A number of different outcomes as identified by ICD-10 codes were analyzed. The main conclusions of the original study were that the risk of *Myocardial infarction* was higher among Apixaban patients in comparison with Dabigatran and Rivaroxaban, that Rivaroxaban patients also had a higher risk of *Pulmonary embolism* than Apixaban patients, and that the risk of *Other bleed* and *Gastrointestinal bleed* was higher among Rivaroxaban patients than for Apixaban and Dabigatran patients.

A causal analysis of this data is frustrated by a number of factors beyond the control of the investigator. For the study period Scotland was covered by 14 health boards each with their own prescribing rules. Economic and other factors may see health boards select particular DOACs as the drug of choice for patients,[3] limiting or eliminating the exposure of patients to the other DOACs. As an example in the Forth Valley health board there were 487 days between the last recorded prescription for Dabigatran and the end of the study (Table 1). Indeed over the course of the study, there was no single continuous time period where all three treatments appeared to be available to all patients, although for each individual health board, there were there were periods of overlap. A further analysis of this type of study would typically involve propensity scoring or IPTW[11] and an IPTW analysis restricted to patients who initiated treatment during these overlapping periods was in agreement with the original study conclusions and provided no evidence for confounding by treatment indication.

In order to further the analysis we consider what additional information exists in data that we can make use of in a different modeling approach. While time to first adverse outcome or event is an important safety measure, any following outcomes also provide information on treatment safety. Not including these recurrent outcomes may reduce the power to detect differences between treatments. Similarly censoring at treatment switching, or discontinuation followed by a treatment restart, also removes recurrent outcomes from the analysis for that patient. For treatments like DOACs, where the drug half-life may be less than a day,[12] a patient initiating a different treatment after a discontinuation could in effect be considered to have started afresh and this data should be included in an analysis. Patients may have many different outcomes over the duration of their treatment. Assessing the relative risks of these outcomes when deciding a treatment regime for a patient requires some sort of combined analysis, which is not readily available from single event outcome models such as Cox regression. The ability to take advantage of relationships that may exist between different outcomes in an analysis should allow for more precise effect estimation and hence reliable decision making. Lack of balance within observational data may be catered for by dividing the population into clusters with similar characteristics, allowing within cluster inferences to be made.

Hierarchical Bayesian models for multiple outcome modeling have been proposed both for clinical trial data and for data mining observational data.[9,10,13-16] These approaches offer the possibility of both borrowing strength between

[3]http://www.ggcmedicines.org.uk/blog/edoxaban-doac-choice-non-valvular-atrial-fibrillat/

the different effects and shrinking nonsignificant effects toward zero, while controlling for multiple comparisons.[17] In the next section, we look to extend some of these clinical trial approaches to allow multiple treatments, with patients grouped into stratified clusters, for use in an observational setting. The use of identified clusters rather than covariates to stratify data has a number of advantages. It reduces the number of variables in the model, particularly when the model is hierarchical, and it allows the use of any number of stratifying approaches, including genetic markers if they exist, without any changes to the model structure. Clusters or groups of patients may be identified by matching, as in case-control studies; however, for rare outcomes where large numbers of clusters would reduce the numbers of outcomes in each cluster to very low levels, unsupervised approaches may be needed to define a limited number of potentially interesting clusters. This type of approach may not be as easily achieved with covariate-based models. In these cases, there may be a level of uncertainty with regard to cluster membership and there is the potential to model this uncertainty and to integrate this into the modeling process. While recent approaches to the analysis of observational data have used regularization methods to allow the estimation of propensity scores from models containing thousands of variables,[18] and applied these methods in large scale studies,[19] the model we present here is complementary to the type of targeted univariate analysis generally used in these approaches. Rather it is designed to analyze multiple related outcomes using defined clusterings of patients and to determine which clusters of patients are more likely to suffer outcomes, while adjusting for multiple comparisons. Outcomes identified as possibly associated with particular treatments and clusters are potential candidates for further targeted analysis.

3 | METHODS

An important consideration when adapting trial methods to observational data is the question of incorporating patient level characteristics into the analysis. In observational data, there is no guarantee of balance between comparator groups and with possibly thousands of patients' data available including individual patient-level parameters and patient/treatment interactions may lead to a model with thousands of parameters. For a Bayesian analysis using a sampling approach, particularly on systems with limited memory, this may make the model computationally intractable. However, the inherent bias in patient data requires a level of patient level input. Some existing approaches include patient level effects in their models but condition or integrate them out, leaving in effect a model containing only treatment effects.[16] Alternatively, stratifying patients into distinct groups, and including this in the model, is an approach that allows the inclusion of some level of patient effects. Stratifying the patients into well-balanced clusters should also provide a level of control for confounding or exploring differences between patient groups. The approach taken here is to include stratification directly in the model hierarchy if required, where the assumption is that while the outcomes in the different strata may be different, there may be a relationship between the different strata that should be included in the model.

The model proposed in this article is a conditional Poisson model based on the approaches in Berry and Berry and Xia et al,[9,13] where the treatment effects are defined as increases in risk relative to a baseline treatment. The related outcomes are modeled following a hierarchical structure with each outcome belonging to a particular grouping of related outcomes. If there are C treatments, H strata, B groupings of outcomes with k_b outcomes in group b, then we model the number of outcomes, $X^{(c)}_{bj,h}$, for the jth outcome in the bth group for treatment c in stratum h by the following:

$$\begin{aligned} X^{(c)}_{bj,h} &\sim \text{Poisson}\left(\lambda^{(c)}_{bj,h} T^{(c)}_{bj,h}\right) \\ T^{(c)}_{bj,h} &= \sum_{i \in \mathcal{R}^{(c)}_{bj,h}} t_{ih} \\ \log \lambda^{(c)}_{bj,h} &= \gamma_{bj,h} + x_{(c)}\theta^{(c)}_{bj,h}, \end{aligned} \tag{1}$$

where $h = 1, \ldots, H, b = 1, \ldots B, j = 1, \ldots k_b, c = 1, \ldots, C$. $x_{(c)}$ is an indicator variable for the treatment, $T^{(c)}_{bj,h}$ is the total time spent under treatment c for all subjects in stratum h, $\lambda^{(c)}_{bj,h}$ is the corresponding underlying rate parameter, $\gamma_{bj,h}$ is the log rate of the baseline treatment, and $\theta^{(c)}_{bj,h}$ may be considered as the increase or decrease in risk relative to the baseline treatment.

As this is a Bayesian model, the individual parameters have prior distributions. Following Berry and Berry,[9] we include the possibility of no difference between the different treatments through a mixture prior including a point-mass at zero

which assigns a positive probability to this possibility:

$$\gamma_{bj,h} \sim N\left(\mu_{\gamma b}, \sigma^2_{\gamma b}\right) \quad \theta^{(c)}_{bj,h} \sim \pi^{(c)}_b \mathrm{I}_{[\theta^{(c)}_{bj,h}=0]} + \left(1-\pi^{(c)}_b\right) N\left(\mu^{(c)}_{\theta b}, (\sigma^{(c)}_{\theta b})^2\right). \tag{2}$$

A three-level hierarchy is defined for the remainder of the model. The full model is given in Appendix A1. The presence of the point-mass term in the distribution of $\theta^{(c)}_{bj,h}$ (2) in effect provides a barrier or hurdle that requires a strong signal (increase or decrease in rate) in order for $\theta^{(c)}_{bj,h}$ to have a high posterior probability of being positive or negative. Removing the point-mass term (setting $\pi^{(c)}_b = 0$ in (2)) tends to result in increases both in detection and error rates when using $\theta^{(c)}_{bj,h}$ as a means of flagging outcomes as being associated with a treatment.[10,20]

The model is an extension of methods proposed for clinical trial data adapted to multiple treatments and stratified data.[10] Individual patient treatment times and outcomes are summed within the different strata, leading to a summary data model. A possible criticism of the model is the potential impact that a small number of patients experiencing a large number of outcomes may have on the model. This is a risk in all summary-level models, but patient stratification has the potential to insulate patients in different clusters from each other while possibly allowing identification of these types of patients if they have similar characteristics. Individual treatments are considered to be independent of each other, they do not borrow strength. The inclusion of relationships between the different treatments requires the incorporation of domain specific knowledge in the model. In the case of the DOAC study in Section 2, this would require knowledge of the pharmacokinetics of the treatments. A strength of Bayesian hierarchical modeling is that this type of knowledge can readily be incorporated into the model,[15] making the model more specific to particular treatment areas. In a similar manner prior knowledge of the known or expected physiological effects for a particular treatment may also be included. The assumption of independence of treatment should allow comparisons with existing study results, with the inclusion of the point-mass term in the model providing a level of error control. The original model proposed by Berry and Berry[9] effectively treats all patients as a single grouping. Including clustering in the model facilitates balanced comparisons for different groups of patients, thus helping to identify differences between groupings and control for confounding. The borrowing of strength between clusters provides a level of robustness to inferences made about differences between clusters in the sense that if there are different effects in different strata then they will need to overcome the effect provided by the modeled relationships. It is also possible to weaken the assumed relationships between clusters by making changes to the model hierarchy. The derivation of the cluster and outcome groupings is not part of the modeling approach presented here. Clustering may be dependent on the level of available patient information, and outcome groupings require knowledge of the behavior of the treatments. A reference implementation of the model is given in the `R` package `bhpm`.[21] The model is fitted by Markov chain Monte Carlo (MCMC) methods, in this case Gibbs sampling under assumptions of conditional independence within the model hierarchy.[22]

We can contrast the approach to the Bayesian self-controlled cases series model (BSCCS), introduced by Shaddox et al,[16] and to the approach of Crooks et al.[15] In the BSCCS model, a drug-era is defined as an interval over which the drugs a patient takes remain constant. In BSCCS, the number of adverse outcomes Y_{ikp} for patient i, for outcome p, in drug-era k is modeled as:

$$\begin{aligned} Y_{ikp} &\sim \text{Poisson}(l_{ikp}\lambda_{ikp}) \\ \log \lambda_{ikp} &= \phi_{ip} + \mathbf{x}^T_{ikp}\boldsymbol{\beta}_p \end{aligned} \tag{3}$$

where ϕ_{ip} are the subject's baseline risks, $\mathbf{x}_{ikp}$ indicates drug exposure, and $\boldsymbol{\beta}_p$ are the log relative risks for each drug with respect to the outcomes p. In BSCCS, the subject baseline effects are conditioned out of the model and we essentially end up with a model which contains only the relative risk effects for each drug. As the model has been conditioned fitting is done using a numerical maximum a posteriori (MAP) estimation approach rather than by sampling. Crooks et al[15] specify an approach using a combination of Bayesian and classical methods for detecting gastrointestinal adverse outcomes of different vaccines. The approach is multilayered, using the reporting odds ratio (ROR)[23] as the method of determining safety signals. The ROR is "propagated" through a number of different classical and Bayesian methods. Initially potential confounders (eg, age/sex/year of event) are identified and a stratified analysis carried out for each combination of vaccine and adverse event. Identified confounders are then used in a logistic regression for the ROR and the estimated log odds ratios and standard errors included in a Bayesian hierarchical analysis and reestimated. This approach to confounding, including the stratified estimates in the Bayesian model, is similar in spirit to the approach above, but the method of

determining signals is different. Both the methods of Shaddox et al and Crooks et al are designed to be used on reporting databases, in the case of Shaddox et al on large-scale claims data, and for Crooks et al on spontaneous reporting databases. While these provide valuable sources of information, one criticism of studies based on this type of data is that they "only contain an indication of a reporter's suspicion of an association rather than a real association."[15] Further analysis is needed to in order to determine if these are real signals. National healthcare records, on the other hand, offer the possibility of a more complete picture of the population under treatment, including accessing patients' prior histories, comorbidities, concomitant medications, laboratory, and, increasingly, genetic data. This data offer the opportunity of performing natural experiments and investigating the performance of treatments in the population as a whole. It is anticipated that, similar to clinical trials, a small number of treatments will be compared. Given the problems associated with performing a causal analysis encountered in Section 2, alternative methods are needed. A novel aspect of the approach here is the application of multivariate methods to data which has typically been analyzed one outcome at a time, while allowing comparisons between different groups of patients, with a level of generality that is not specific to one particular treatment area.

4 | RESULTS

4.1 | Simulation study

The methods are illustrated and evaluated by a synthetic study of 14 000 patients, split into 10 different clusters (Cluster1-Cluster10) and treated with one of four different drugs (Drug1-Drug4). The purpose of the simulation is to assess the detection and error rates of the methods with patients being randomly assigned to clusters with probabilities given in Table B3. There are different numbers of patients and treatment allocations in each cluster and the average time under treatment for each patient also varies between the clusters. A single treatment, Drug1, is chosen as the baseline and we are interested in detecting changes in the relative risks between the other treatments and the baseline treatment. Nine different outcomes (Outcome1-Outcome9) are included in the study, divided into three groups (Table 2). The underlying baseline outcome rates for the simulation are sampled from a normal distribution with mean rate of 0.001 per unit time, and SD 0.0001 (simulated negative rates are set to 0.001). All outcomes in group Group2 have increased rates over all the clusters for treatment Drug2, and Outcome7 and Outcome8 in group Group3 have increased rates for treatment Drug3 in cluster Cluster9 only. The details are given in Table 3. Thousand simulations in total were run. The full details of the study are given in Appendix B1. The simulation explicitly caters for the presence of different effects in different clusters. Two different types of analyses are performed: one with a separate model fitted to each individual cluster, and one where all clusters are included in a single hierarchy. Models with and without the point-mass ($\pi_b^{(c)} = 0$ in (2)) are fitted in all cases.

Although the models may be considered exploratory, in order to provide some assessment of their performance with regard both to the detection of raised outcome rates and the error rates, some flagging mechanism must be used. As the methods are Bayesian, we use the posterior probability of an increase or decrease in the event rate relative to the baseline

TABLE 2 Simulation study outcomes

Group	Outcome
Group1	Outcome1, Outcome2
Group2	Outcome3, Outcome4, Outcome5
Group3	Outcome6, Outcome7, Outcome8, Outcome9

TABLE 3 Outcomes with increased rates

Treatment	Outcome	% increase in rate	Cluster
Drug2	Outcome3	100	All clusters[a]
Drug2	Outcome4	50	All clusters[a]
Drug2	Outcome5	1	All clusters[a]
Drug3	Outcome7	100	Cluster9
Drug3	Outcome8	10	Cluster9

[a]Rate raised for all clusters in the simulation.

TABLE 4 Results of the simulation study

Method	Clustered analysis[a]	Correct[b]	Incorrect[c]	Missed[d]	Raised rates[e]	Baseline comparisons[f]
Point-mass	Yes	20 183	7	11 817	32 000	270 000
Point-mass	No	18 391	67	13 609	32 000	270 000
No point-mass	Yes	21 669	6792	10 331	32 000	270 000
No point-mass	No	20 965	10 105	11 035	32 000	270 000

[a]The value Yes means a single model (1) was fitted to the data. The value No means that individual models (1) were fitted for each cluster.
[b]The total number of outcome, treatment, cluster combinations with raised rates compared with the baseline treatment that were correctly identified by the model as having a raised rate.
[c]The total number of outcome, treatment, cluster combinations that were incorrectly identified by the model as having a raised rate compared with the baseline.
[d]The total number of outcome, treatment, cluster combinations with raised rates that were not identified by the model as having a raised rate compared with the baseline
[e]The total number of outcome, treatment cluster combinations with raised rates compared with the baseline in the simulation study. From Table 3, Outcome3-Outcome5 are each raised in 10 clusters for Drug2, Outcome7, and Outcome8 are raised in one cluster each for Drug3. This gives 32 combinations in total with raised rates per simulation.
[f]The total number of outcome, treatment, cluster combination comparisons with the base line treatment Drug1. There are three comparison treatments (Drug2-Drug4), each with nine outcomes over 10 clusters, giving 270 comparisons with the baseline treatment per simulation.

treatment ($\theta^{(c)}_{bj,h}$) as a method of determining if an outcome is associated with a treatment (relative to the baseline).[9] For the purposes of this study we declare that an outcome is associated with a treatment if the posterior probability of an increase/decrease in rate compared with the baseline is greater than 90% for point-mass models and 95% for models without the point-mass. Previous simulation studies for similar models have shown that the 90% cut-off for the point-mass model allows the detection of outcomes associated with treatment, without inflating the misclassification rate (incorrectly flagging events as associated with treatment), particularly for higher rate adverse events.[10] For low treatment and control differences the effect of the point-mass is felt most strongly, and lowering the threshold below 90% does not lead to a large increase in the numbers detected. For the non-point-mass models, a threshold of 95% has proven to be similarly suitable for simulation studies.[10] The situation for calibrating model thresholds in real as opposed to simulated data is necessarily different with thresholds as high as 99% being suggested.[15] A more theoretical method suggested by Chen et al[24] is to use a decision theoretic approach to minimize the loss of misclassifying an outcome as the means of determining threshold values.

The results in terms of detection and error counts are given in Table 4. We can see that in terms of balancing between numbers of outcomes correctly detected and errors made the clustered point-mass model could be considered to have performed best overall. It correctly identified 20 183 out 32 000 possible combinations (63%) as having raised treatment rates, and misclassified only seven outcomes as being associated with treatment. While the clustered non-point-mass model has correctly identified more outcome combinations, 21 669 or 68%, it has misclassified 6792 outcomes combinations. Comparing the clustered models to their nonclustered equivalents, we can see there is both the evidence of borrowing strength between related outcomes (increase in the numbers correctly detected) and the shrinkage of nonsignificant effects toward zero (decrease in the misclassifications) in the clustered model.

Care must be taken when generalizing from simulations. Previous results with these types of model indicate that for very low outcome rates or short study periods, the non-point-mass models may perform better than their point-mass equivalent in terms of outcome detection.[10] In these cases, the point-mass models may detect few outcomes with raised rates. Here the effect of the point-mass mitigates against detection. However even for shorter durations, the error rate in terms of incorrectly flagged outcomes remains much higher than for non-point-mass models.[10,20] A number of additional simulation scenarios considered in the supplementary materials illustrate these model features. For simulations where the difference in rate is small (Section S.5), the point-mass model struggles to identify outcomes with increased rates. Outcomes with low rate increase are unable to overcome the effect of the point-mass, whereas the non-point-mass models correctly detect many more of the outcomes with increased rates, with the rate of misclassification of outcomes as associated with treatment, being comparable with the other simulations. For simulations with large differences in treatment outcome rates (Section S.4), the situation is very different. The numbers of outcomes with raised rates that are correctly detected by the clustered point-mass model exceeds that detected by the non-clustered point-mass model

TABLE 5 Outcomes included in the study

Outcome description	Grouping	Grouping description
Ischemic stroke	I00-I99	Circulatory system
Hemorrhagic stroke	I00-I99	Circulatory system
Systemic embolism	I00-I99	Circulatory system
Pulmonary embolism	I00-I99	Circulatory system
Myocardial infarction	I00-I99	Circulatory system
Transient ischemic attack	G00-G99	Nervous system
Gastrointestinal bleed	Bleed	Bleeds
Other bleed	Bleed	Bleeds
Other ADR	Other ADR	Adverse drug reaction

and is comparable with the numbers detected by the clustered non-point-mass model, with much better control of the misclassification rate.

4.2 | Direct oral anticoagulant Scotland study (2011-2015)

The clustered point-mass model is applied to the data from the comparative safety and effectiveness study of direct oral anticoagulants (DOACs) in Scotland.[2] The outcomes analyzed are given in Table 5, together with the groupings used for the Bayesian analysis. The choice of outcomes corresponds to those used in the original study and the same assumptions were followed with regard to treatment discontinuation and treatment switching. Counts of all the relevant outcomes, which occurred to any patient while under any treatment over the duration of the study, were included. The outcomes are aggregates of different ICD-10 codes, which relate to similar medical incidents.[2] The choice of outcomes and groupings highlights a number of issues with aggregating ICD-10 codes and determining groupings when there is no single standard available. The groupings used here are guided by ICD-10 code groupings, for example, *Circulatory system* covers I00-I99. However, the aggregated outcome *Other bleed*, used in the original study, also contains the ICD-10 code I62, so we could plausibly include some of the outcomes in the *Circulatory system* grouping rather than the *Bleeds* grouping as we have chosen.

Due to the relatively small numbers of outcomes, the number of patient stratifications are limited. Two analyses were performed. One where patients are grouped in a single cluster and one where patients were stratified into a small number of clusters based on age category in years (0-64, 65+), sex (Male/Female), and number of concomitant medications (0-4, 5+), giving eight clusters in total (Table 6). Cluster size is a problem regardless of the type of analysis performed. While age and sex are often included in analyses as a priori confounders,[15] concomitant medicines may be considered a proxy for how ill a patient is. The stratifying variables are chosen to demonstrate the method and to allow a comparison with the DOAC Scotland study results. The identification of potential confounders is not part of this analysis. As in the original study, Rivaroxaban was chosen to be the baseline treatment. The distributions of the treatment exposures across the clusters is shown in Figure 1. We can see similar patterns of time under treatment in each cluster but with larger exposure times across all treatments in the older groups with 5+ concomitant medications. We may expect that rare outcomes will show up more in the groups who have been under treatment for the longest times. Figure 2 shows the summary data and 90% posterior intervals for the increase or decrease in outcome rate of Apixaban and Dabigatran compared with Rivaroxaban for the single-cluster analysis. Outcomes which exceed the 90% posterior probability of an increase or decrease in rate compared with the baseline (Rivaroxaban) are an increase in *Myocardial infarction* for Apixaban, a decrease in *Other bleed* for both Apixaban and Dabigatran, and a decrease in *Pulmonary embolism* for Dabigatran (Table 7). The aggregation of the overall study population into a single cluster may hide differences in treatment behavior, which are cluster specific. Of interest is investigating if the outcomes flagged in Table 7 in the single cluster analysis are flagged in each of the individual clusters as defined in Table 6 and if there are any additional outcomes which are not flagged in the single cluster analysis, but which have a high posterior probability of an increase or decrease in rate in any of the clusters in the multicluster model.

TABLE 6 Cluster characteristics

Cluster name	Sex	Age	Concomitant Medications	Size
M/0-64/0-4	M	0-64	0-4	634
M/0-64/5+	M	0-64	5+	1382
M/65+/0-4	M	65+	0-4	673
M/65+/5+	M	65+	5+	5367
F/0-64/0-4	F	0-64	0-4	133
F/0-64/5+	F	0-64	5+	596
F/65+/0-4	F	65+	0-4	455
F/65+/5+	F	65+	5+	5548

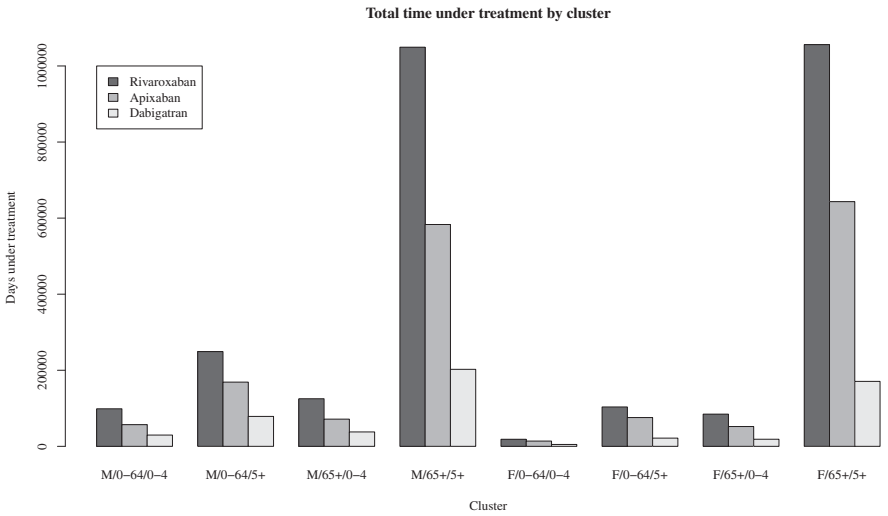

FIGURE 1 Total time under treatment in days for each cluster

When the models were fitted to the multicluster data 10 combinations of outcome, treatment, and cluster were flagged as exceeding the 90% posterior probability of an increase or decrease in rate compared with the baseline. The single cluster results flagged in Table 7 accounted for seven of these. These outcomes generally had high posterior probabilities in many of the clusters but not always above the 90% threshold and in some clusters there was no evidence of a difference. Taking *Myocardial infarction* as an example, there was little or no evidence in clusters M/0-64/0-4, M/65+/0-4, and F/0-64/0-4 of an increase in rate compared with Rivaroxaban (Figure 3). However clusters M/0-64/0-4, M/65+/0-4, and F/0-64/0-4 are approximately nine times smaller than the largest clusters (M/65+/5+ and F/65+/5+), with lower overall exposure times and with correspondingly fewer events. The remaining three combinations for the outcomes *Pulmonary embolism*, *Other ADR*, and *Ischemic stroke* were flagged only in certain clusters (Table 8). Here there is some evidence of a decrease in *Pulmonary embolism* and *Other ADR* for females over the age of 65 years with five or more concomitant medicines when using Apixaban compared with Rivaroxaban, and also a decrease in *Ischemic stroke* for males aged under 65 years with five or more concomitant medications when using Dabigatran compared with Rivaroxaban. However, the posterior probabilities in the other clusters indicate no appreciable differences between the treatments.

4.2.1 | Discussion

The Bayesian model suggests that compared with Rivaroxaban, there is an increase in *Myocardial infarction* when being treated by Apixaban, and that there are lower rates of *Other bleed* for both Apixaban and Dabigatran and lower rates of *Pulmonary embolism* for Dabigatran. There are also indications of possible decreases of *Pulmonary embolism* and *Other ADR* for Apixaban and *Ischemic stroke* for Dabigatran, compared with Rivaroxaban for some groups.

All outcomes

Treatment	Grouping	Outcome	$P(\theta < 0)$	$P(\theta = 0)$	$P(\theta > 0)$
Apixaban	I00–I99	Myocardial infarction	0.000	0.000	1.000
Apixaban	I00–I99	Haemorrhagic stroke	0.020	0.420	0.560
Apixaban	I00–I99	Systemic embolism	0.060	0.501	0.439
Apixaban	I00–I99	Ischemic stroke	0.045	0.777	0.178
Dabigatran	G00–G99	Transient ischaemic attack	0.058	0.833	0.109
Apixaban	G00–G99	Transient ischaemic attack	0.077	0.823	0.100
Dabigatran	I00–I99	Myocardial infarction	0.100	0.859	0.041
Apixaban	I00–I99	Pulmonary embolism	0.447	0.517	0.035
Dabigatran	OTHER_ADR	Other ADR	0.305	0.663	0.032
Apixaban	BLEED	Gastrointestinal bleed	0.302	0.667	0.031
Dabigatran	I00–I99	Systemic embolism	0.455	0.516	0.030
Dabigatran	I00–I99	Ischemic stroke	0.115	0.858	0.027
Dabigatran	BLEED	Gastrointestinal bleed	0.304	0.672	0.024
Dabigatran	I00–I99	Haemorrhagic stroke	0.606	0.384	0.010
Apixaban	OTHER_ADR	Other ADR	0.794	0.200	0.006
Dabigatran	I00–I99	Pulmonary embolism	0.957	0.042	0.001
Apixaban	BLEED	Other bleed	0.943	0.056	0.000
Dabigatran	BLEED	Other bleed	1.000	0.000	0.000

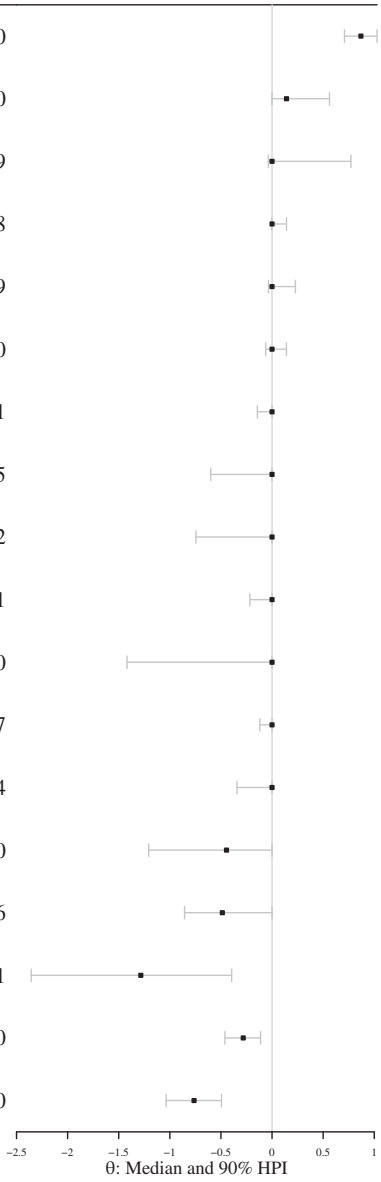

FIGURE 2 Posterior summaries of the increase/decrease in rate (θ) for all outcomes compared with baseline treatment Rivaroxaban. All posterior probabilities rounded to three decimal places

While the original study looked at modeling single outcomes up to occurrence or a censoring point,[2] the analysis in this section took into account all outcomes of interest occurring under treatment, including treatment switches. Even with this difference in approach the results in this section are widely in agreement with those reported in the original study.[2] Both analyses identified that the risk of *Myocardial infarction* was higher among Apixaban patients in comparison with Rivaroxaban and Dabigatran, and the risk of *Other bleeds* was higher among Rivaroxaban patients than Apixaban and Dabigatran. For *Pulmonary embolism*, the point-mass model did indicate the possibility of a decrease in rate for Apixaban but this was only flagged in the cluster F/65+/5+, for the nonclustered model we can see that overall there is some indication of reduced rates for Apixaban (Figure 2). *Pulmonary embolism* is however flagged as having a reduced rate for Dabigatran. For *Gastrointestinal bleed*, which had an increased risk for Rivaroxaban compared with Apixaban and Dabigatran in the original analysis, there were also indications of increased rate in the Bayesian analysis but none that were flagged at the 90% posterior probability level in any of the clusters. We can see this in Figure 4.

TABLE 7 Outcomes with increased/decreased rate compared with Rivaroxaban (greater than 90% posterior probability)

Outcome	Grouping	Treatment	Increase/decrease
Myocardial infarction	Circulatory system	Apixaban	Increase
Other bleed	Bleeds	Apixaban	Decrease
Other bleed	Bleeds	Dabigatran	Decrease
Pulmonary embolism	Circulatory system	Dabigatran	Decrease

Myocardial infarction (Apixaban)

Cluster	$P(\theta < 0)$	$P(\theta = 0)$	$P(\theta > 0)$
M/0–64/5+	0.000	0.000	1.000
F/65+/5+	0.000	0.000	1.000
M/65+/5+	0.000	0.001	0.999
F/0–64/5+	0.003	0.028	0.969
F/65+/0–4	0.075	0.256	0.669
F/0–64/0–4	0.407	0.320	0.273
M/0–64/0–4	0.460	0.305	0.235
M/65+/0–4	0.680	0.237	0.083

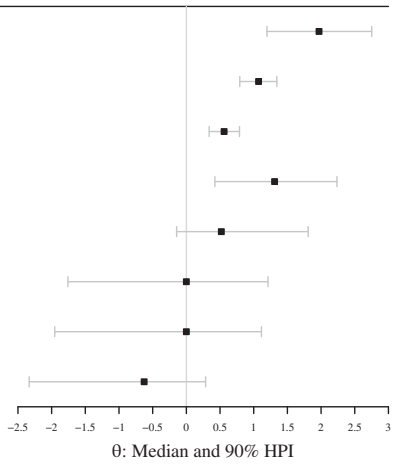

FIGURE 3 Posterior summaries of the increase/decrease in *Myocardial infarction* rates for Apixaban compared with baseline treatment Rivaroxaban). All posterior probabilities rounded to three decimal places

TABLE 8 Other outcomes with decreased rates compared with Rivaroxaban by cluster

Outcome	Grouping	Treatment	Cluster
Pulmonary embolism	Circulatory system	Apixaban	F/65+/5+
Other ADR	Adverse Drug Reactions	Apixaban	F/65+/5+
Ischemic stroke	Circulatory system	Dabigatran	M/0-64/5+

Including concomitant medications as a stratifying variable poses a number of interesting questions. While it may be considered a proxy for how ill a patient is, including only counts of medications leaves out the additional information regarding what the medications are and how they may affect the overall balance of the clusters. The actual effect of the medicines is unquantifiable in this analysis. Referring to Figure 3 where a number of clusters have no indication of an increase in *Myocardial infarction*, we know from Figure 1 that some of these clusters have lower overall exposure times. A question here is whether the differences between different clusters is due to the low event rates and exposure times or due to some other reason, for example, the unknown effect of the concomitant medications. A clustering approach taking into account the types of medicines that patients were prescribed rather than just the counts may be a better way of stratifying the data. In the general case, as data accumulates over time, it should be possible to see the emerging patterns of events more clearly. While confounding cannot be ruled out and it cannot be guaranteed that the clusters are well balanced, this particular analysis was designed to be comparable with the original Scotland DOAC study.[2] The ability to plot and assess the outcomes using posterior probabilities allows an assessment to be made of differences between different clusters and which of these differences are worthy of further investigation. In general, the inclusion of clusters into the model does allow for an analysis of broadly similar groups, limited only by the number of patients in the study, the numbers of outcomes, and the information that is available.

Gastrointestinal bleed (Apixaban/Dabigatran)

Trt.Grp	Cluster	$P(\theta < 0)$	$P(\theta = 0)$	$P(\theta > 0)$
Apixaban	M/65+/5+	0.250	0.714	0.035
Apixaban	M/65+/0–4	0.794	0.187	0.018
Apixaban	M/0–64/5+	0.113	0.633	0.254
Apixaban	M/0–64/0–4	0.520	0.421	0.059
Apixaban	F/65+/5+	0.167	0.780	0.053
Apixaban	F/65+/0–4	0.740	0.241	0.019
Apixaban	F/0–64/5+	0.511	0.439	0.049
Apixaban	F/0–64/0–4	0.552	0.385	0.064
Dabigatran	M/65+/5+	0.150	0.779	0.071
Dabigatran	M/65+/0–4	0.722	0.254	0.023
Dabigatran	M/0–64/5+	0.399	0.541	0.060
Dabigatran	M/0–64/0–4	0.282	0.612	0.105
Dabigatran	F/65+/5+	0.264	0.695	0.041
Dabigatran	F/65+/0–4	0.553	0.411	0.036
Dabigatran	F/0–64/5+	0.692	0.281	0.027
Dabigatran	F/0–64/0–4	0.552	0.397	0.051

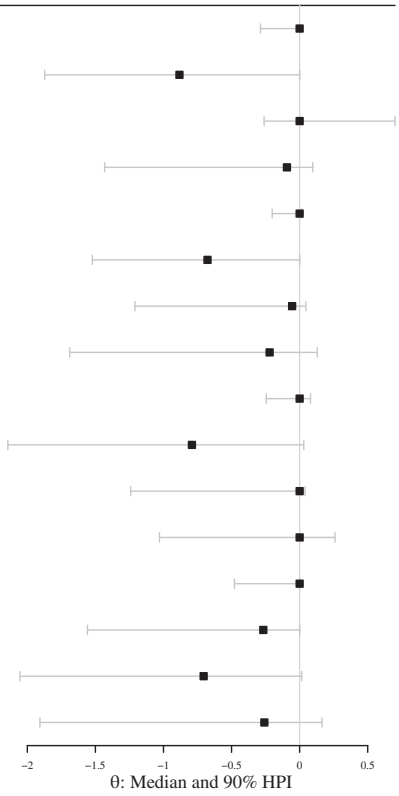

FIGURE 4 Posterior summaries of the increase/decrease in *Gastrointestinal bleed* infarction rates for Apixaban compared with baseline treatment Rivaroxaban). All posterior probabilities rounded to three decimal places

5 | DISCUSSION

The availability of large scale databases containing the health records of thousands of patients, each record containing multiple variables, provides an opportunity and a requirement to move beyond a single variable analysis and develop methods capable of analyzing multiple outcomes, providing the possibility of moving toward a more precise approach to medicine delivery. In order to harness the data, a number of challenges must be met. Balance between comparator groups is not guaranteed in observational data, and the application of propensity scoring methods for a fully causal analysis may not be possible. Multivariate methods require the assumption that relationships exist between different outcomes, and the outcomes themselves may be rare. Hierarchical relationships among outcomes, such as those defined by medical dictionaries and used in clinical trials, are particularly useful and suitable for statistical modeling. However, a similar well-defined structure does not exist for the type of clinical data routinely recorded for patients in the general population. The lack of balance between comparator groups may be addressed by matching,[6] and the possibility of the existence of genetic markers as a potential method for stratifying groups has the ability to address this issue more fully.[7]

Hierarchical Bayesian methods are a suitable approach for analyzing this type of data. Hierarchies of outcomes may be easily incorporated in models, and posterior probabilities provide a method for assessing outcome occurrence. Bayesian methods are also well suited to handling the analysis of constantly accumulating data, such as healthcare records. These methods are well understood and have the advantage of being relatively easy to explain to clinicians. The methods outlined in this article take a summary approach to data modeling where patients' times under treatment are combined within different clusters to provide an analysis of the differences between different treatments within these clusters. The inclusion of the point-mass term has the effect of requiring a strong signal in order for an outcome to be a candidate for being associated with a treatment. The model is flexible with regard to choice of clusters and hierarchies. The longitudinal analysis of outcomes may be affected by a number of factors. It is not possible to guarantee patient adherence and

there may also be gaps in the prescription record (data quality). It is also not easily possible to determine different periods of drug action that would successfully cover all treatments. For some treatments, short discontinuations may be important. This is the case for drugs such as direct oral anticoagulants where the half-life may be less than a day.[12] For other treatments, where the risk of cumulative damage from the treatment is large, the patient may continue to be at risk after discontinuation, and the outcome rate itself may be variable.

The use of machine learning techniques to address complex healthcare problems is growing[25] and in particular, the use of hierarchical models and modeling of relationships between outcomes is becoming more common in healthcare data analysis.[9,13,16] Modeling outcomes simultaneously provides a number of advantages over single variable analyses including multiple comparison control, the borrowing of strength between effects, and the shrinkage of nonsignificant effects toward zero. However beyond the modeling of outcomes real clinical decisions need to made. The stratification of a population into a set of balanced clusters, and the inclusion of this in the model, brings forward the possibility of making clinical decisions based on treatment comparisons within these clusters. However, the model alone is unable to provide this and would require an associated decision-making procedure. Looking beyond the models presented here, which assumes that patients are stratified and that groupings are fixed, there is the possibility of moving to a more integrated Bayesian approach to this type of healthcare analysis, with the possibility of the inclusion in the model of treatment relationships and uncertainty regarding cluster membership, outcome groupings, and treatment allocation,[26] allowing an extendable probabilistic framework.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the support of the electronic Data Research and Innovation Service (eDRIS) Team (National Services Scotland) for their involvement in obtaining approvals, provisioning and linking data and the use of the secure analytical platform within the National Safe Haven. This research was supported by Health Data Research (HDR) (UK) @ Scotland, Medical Research Council (MRC) award reference MR/S003967/1. HDR (UK) is an independent nonprofit organization bringing together 22 research institutes across the UK, supported by 10 funders: The British Heart Foundation, Chief Scientist Office (Scotland), Engineering and Physical Sciences Research Council (EPSRC), Economic and Social Research Council (ESRC), Health and Care Research (Wales), Health and Social Care Research and Development Division (N. Ireland), The Medical Research Council (MRC), The National Institute for Health Research (NIHR), Wellcome, UK Research and Innovation.

The DOAC Scotland Study (2011-2015) data are not publicly available due to privacy restrictions. The simulation data and the code used to support the finding of this study are publicly available.[2021]

CONFLICT OF INTEREST

R.C. was employed by Roche Pharmaceutical (UK) in June-July 2019.

ORCID

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

REFERENCES

1. Friedman LM, Furberg CD, DeMets DL. *Fundamentals of Clinical Trials*. New York, NY: Springer; 2010.
2. Mueller T, Alvarez-Madrazo S, Robertson C, Wu O, Marion B. Comparative safety and effectiveness of direct oral anticoagulants in patients with atrial fibrillation in clinical practice in Scotland. *Br J Clin Pharmacol*. 2019;85(2):422-431. https://doi.org/10.1111/bcp.13814.
3. Gould AL. Accounting for multiplicity in the evaluation of "signals" obtained by data mining from spontaneous report adverse event databases. *Biom J*. 2007;49(1):151-165. https://doi.org/10.1002/bimj.200610296.
4. DuMouchel W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *Am Stat*. 1999;53(3):177-190. https://doi.org/10.2307/2686093.
5. Bate A, Lindquist M, Edwards IR, et al. A Bayesian neural network method for adverse drug reaction signal generation. *Eur J Clin Pharmacol*. 1998;54(4):315-321. https://doi.org/10.1007/s002280050466.
6. MA H, Robins JM. *Causal Inference*. Boca Raton, FL: Chapman & Hall/CRC; 2018.
7. Burgess S, Thompson SG. *Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation*. Boca Raton, FL: CRC Press; 2015.
8. Chuang-Stein C, Mohberg NR, Musselman DM. Organization and analysis of safety data using a multivariate approach. *Stat Med*. 1992;11(8):1075-1089. https://doi.org/10.1002/sim.4780110809.
9. 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.
10. Carragher R. Detection of safety signals in randomised controlled trials (PhD thesis). University of Strathclyde; 2017.

11. Larsen TB, Skjøth F, Nielsen PB, Kjældgaard JN, Lip GYH. Comparative effectiveness and safety of non-vitamin K antagonist oral anticoagulants and warfarin in patients with atrial fibrillation: propensity weighted nationwide cohort study. *BMJ*. 2016;353. https://doi.org/10.1136/bmj.i3189.
12. Ieko M, Naitoh S, Yoshida M, Takahashi N. Profiles of direct oral anticoagulants and clinical usage - dosage and dose regimen differences. *J Intens Care*. 2016;4. https://doi.org/10.1186/s40560-016-0144-5.
13. Amy XH, 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.
14. DuMouchel W. Multivariate Bayesian logistic regression for analysis of clinical study safety issues. *Stat Sci*. 2012;27(3):319-339.
15. 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.
16. Shaddox TR, Ryan PB, Schuemie MJ, Madigan D, Suchard MA. Hierarchical models for multiple, rare outcomes using massive observational healthcare databases. *Stat Anal Data Mining ASA Data Sci J*. 2016;9(4):260-268. https://doi.org/10.1002/sam.11324.
17. Gelman A, Hill J, Yajima M. Why we (usually) don't have to worry about multiple comparisons. *J Res Edu Effect*. 2012;5(2):189-211. https://doi.org/10.1080/19345747.2011.618213.
18. Schuemie MJ, Cepede M, Soledad SMA, et al. How confident are we about observational findings in healthcare: a benchmark study. *Harvard Data Sci Rev*. 2019;2(1). https://doi.org/10.1162/99608f92.147cc28e.
19. Suchard MA, Schuemie MJ, Krumholz HM, et al. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. *Lancet*. 2019;394(10211):1816-1826. https://doi.org/10.1016/S0140-6736(19)32317-7.
20. Carragher R. Supplementary material: a Bayesian hierarchical approach for multiple outcomes in routinely collected healthcare. *Data Simulat Stud*. 2019. https://doi.org/10.5281/zenodo.3250871.
21. Carragher R. bhpm: Bayesian hierarchical possion models for mulitple grouped outcomes with *Clustering*. 2019. https://doi.org/10.0.20.161/zenodo.3246415.
22. Robert CP, Casella G. *Monte Carlo Statistical Methods*. New York, NY: Springer; 1999.
23. Rothman KJ, Lanes S, Sacks ST. The reporting odds ratio and its advantages over the proportional reporting ratio. *Pharmacoepidemiol Drug Saf*. 2004;13(8):519-523. https://doi.org/10.1002/pds.1001.
24. 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.
25. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf*. 2019;28(3):231-237. https://doi.org/10.1136/bmjqs-2018-008370.
26. McCandless LC, Gustafson P, Austin PC. Bayesian propensity score analysis for observational data. *Stat Med*. 2009;28(1):94-112. https://doi.org/10.1002/sim.3460.
27. Lunn D, Jackson C, Best N, Thomas A, Spiegelhalter D. The BUGS Book: A Practical Introduction to Bayesian Analysis. . Boca Raton, FL: Chapman & Hall/CRC Texts in Statistical ScienceTaylor & Francis; 2012.

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, Mueller T, Bennie M, Robertson C. A Bayesian hierarchical approach for multiple outcomes in routinely collected healthcare data. *Statistics in Medicine*. 2020;1–16. https://doi.org/10.1002/sim.8563

APPENDIX A. MODEL DEFINITIONS

There are C treatments, H strata, B groupings of outcomes with k_b outcomes in group b, and the number of outcomes for the jth outcome in the bth group for treatment c in stratum h is $X^{(c)}_{bj,h}$. The model is:

$$X^{(c)}_{bj,h} \sim \text{Poisson}\left(\lambda^{(c)}_{bj,h} T^{(c)}_{bj,h}\right)$$

$$T^{(c)}_{bj,h} = \sum_{i \in \mathcal{R}^{(c)}_{bj,h}} t_{ih}$$

$$\log \lambda^{(c)}_{bj,h} = \gamma_{bj,h} + x_{(c)} \theta^{(c)}_{bj,h}$$

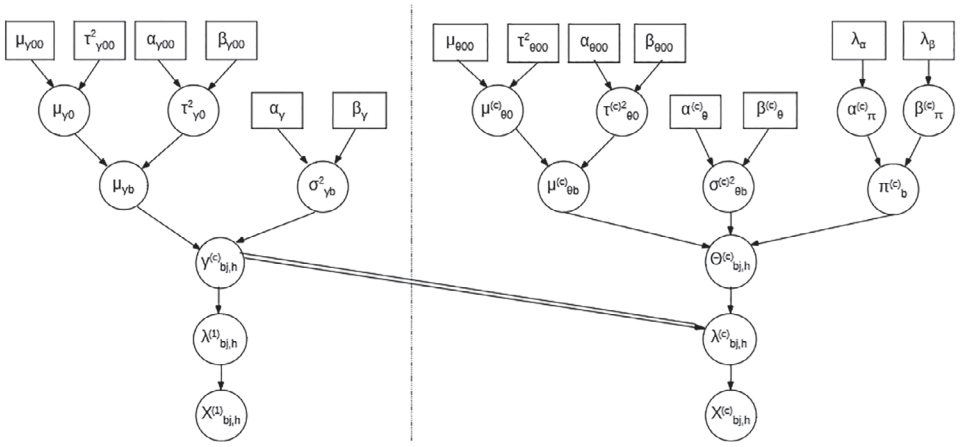

FIGURE A1 Directed acyclic graph for the model

$$h = 1, \ldots, H, \quad b = 1, \ldots B_h, \quad j = 1, \ldots k_{bh}$$
$$c = 1, \ldots, C; \quad x_{(c)} = 1;\ x_{(i)} = 0, i \neq c. \tag{A1}$$

The priors for the model parameters and hyperparameters are given in Equations (A2)-(A4). As this is a three-level hierarchical model we have three levels of priors:

$$\gamma_{bj,h} \sim \mathrm{N}(\mu_{\gamma b}, \sigma^2_{\gamma b}) \quad \theta^{(c)}_{bj,h} \sim \pi^{(c)}_b \mathrm{I}_{[\theta^{(c)}_{bj,h}=0]} + (1-\pi^{(c)}_b)\mathrm{N}(\mu^{(c)}_{\theta b}, (\sigma^{(c)}_{\theta b})^2) \tag{A2}$$
$$\mu_{\gamma b} \sim \mathrm{N}(\mu_{\gamma 0}, \tau^2_{\gamma 0}) \quad \mu^{(c)}_{\theta b} \sim \mathrm{N}(\mu^{(c)}_{\theta 0}, (\tau^{(c)}_{\theta 0})^2)$$
$$\sigma^2_{\gamma b} \sim \mathrm{IG}(\alpha_\gamma, \beta_\gamma) \quad \left(\sigma^{(c)}_{\theta b}\right)^2 \sim \mathrm{IG}(\alpha_\theta, \beta_\theta)$$
$$\pi^{(c)}_b \sim \mathrm{Beta}(\alpha^{(c)}_\pi, \beta^{(c)}_\pi) \tag{A2}$$

$$\mu_{\gamma 0} \sim \mathrm{N}(\mu_{\gamma 00}, \tau^2_{\gamma 00}) \quad \mu^{(c)}_{\theta 0} \sim \mathrm{N}(\mu_{\theta 00}, \tau^2_{\theta 00})$$
$$\tau^2_{\gamma 0} \sim \mathrm{IG}(\alpha_{\gamma 00}, \beta_{\gamma 00}) \quad \left(\tau^{(c)}_{\theta 0}\right)^2 \sim \mathrm{IG}(\alpha_{\theta 00}, \beta_{\theta 00})$$
$$\alpha^{(c)}_\pi \sim \mathrm{M}(\lambda_\alpha)\mathrm{I}(\alpha^{(c)}_\pi > 1) \quad \beta^{(c)}_\pi \sim \mathrm{M}(\lambda_\beta)\mathrm{I}(\beta^{(c)}_\pi > 1), \tag{A4}$$

where I(.) is the indicator function, N is the normal distribution, β is the beta distribution, IG is the inverse-gamma distribution, and M is the exponential distribution.

The following model hyperparameters all have common values over the intervals based on the values in Berry and Berry:[9]

$$\mu_{\gamma 00} = 0, \tau^2_{\gamma 00} = 10, \quad \alpha_\gamma = 3, \quad \beta_\gamma = 1, \quad \alpha_{\gamma 00} = 3, \quad \beta_{\gamma 00} = 1, \quad \lambda_\alpha = 1$$
$$\mu_{\theta 00} = 0, \quad \tau^2_{\theta 00} = 10, \quad \alpha_\theta = 3, \beta_\theta = 1, \quad \alpha_{\theta 00} = 3, \quad \beta_{\theta 00} = 1, \lambda_\beta = 1. \tag{A5}$$

Assuming conditional independence of the parameters the models are fitted using Markov chain Monte Carlo (Gibbs) sampling. The graph of the model is shown in Figure A1.[27]

APPENDIX B. SIMULATION STUDY

The outcomes, their groupings, and the increases in rates for outcomes associated with treatments are presented in Tables 2 and 3. The general simulation parameters are given in Table B1. For each simulation, the underlying baseline rates for the outcomes are sampled from a normal distribution with mean μ and standard deviation σ, with negative rates set to μ. Patients are randomly assigned to the clusters with the probabilities in Table B2. The average treatment time in each cluster was generated from a normal distribution with mean μ_C and standard deviation σ_C. The values used in the simulation are given in Table B2. The average time a patient in each cluster remains under treatment is generated from

TABLE B1 General simulation parameters

Parameter	Description	Value
N	Number of treatments	4
C	Number of clusters	10
N_C	Number of outcomes	9
N_g	Number of outcome groups	3
N_p	Number of patients per simulation	14 000
N_S	Number of simulations	1000
μ	Baseline outcome mean rate	0.001
σ	Baseline outcome standard deviation	0.0001
μ_C	Cluster mean treatment time	339
σ_C	Cluster treatment time standard deviation	150
σ_p	Patient treatment time standard deviation	10

TABLE B2 Cluster simulation parameters

Cluster	Patient assignment probability	Mean treatment duration
Clusters1	0.025	557
Clusters2	0.05	444
Clusters3	0.05	275
Clusters4	0.1	499
Clusters5	0.1	533
Clusters6	0.1	460
Clusters7	0.1	312
Clusters8	0.1	167
Clusters9	0.175	540
Clusters10	0.2	287

TABLE B3 Treatment assignment probabilities

Treatment	Cluster	Patient assignment probability
Drug1	Cluster1-Cluster5	0.2
Drug2	Cluster1-Cluster5	0.2
Drug3	Cluster1-Cluster5	0.2
Drug4	Cluster1-Cluster5	0.4
Drug1	Cluster6-Cluster10	0.25
Drug2	Cluster6-Cluster10	0.25
Drug3	Cluster6-Cluster10	0.25
Drug4	Cluster6-Cluster10	0.25

a normal distribution with mean treatment duration given in Table B2 and standard deviation σ_p. Patients are randomly assigned treatments with probabilities given in Table B3. Outcomes are generated for each patient by a Poisson process with parameter given by the outcome rate multiplied by the treatment duration. The data were summarized by treatment and cluster for input to the model.