HAL open science

A Bayesian non-inferiority approach using experts' margin elicitation – application to the monitoring of safety events

Camille Aupiais, Corinne Alberti, Thomas Schmitz, Olivier Baud, Moreno Ursino, Sarah Zohar

To cite this version:

Camille Aupiais, Corinne Alberti, Thomas Schmitz, Olivier Baud, Moreno Ursino, et al.. A Bayesian non-inferiority approach using experts' margin elicitation – application to the monitoring of safety events. BMC Medical Research Methodology, 2019, 19 (1), pp.187. 10.1186/s12874-019-0826-5 . inserm-02456619

HAL Id: inserm-02456619

https://inserm.hal.science/inserm-02456619v1

Submitted on 27 Jan 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Aupiais *et al. BMC Medical Research Methodology* (2019) 19:187
https://doi.org/10.1186/s12874-019-0826-5

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

A Bayesian non-inferiority approach using experts' margin elicitation – application to the monitoring of safety events

Camille Aupiais[1,2,3,4*], Corinne Alberti[2,3,4], Thomas Schmitz[2,5,6], Olivier Baud[7,8,9], Moreno Ursino[1,4,10†] and Sarah Zohar[1†]

Abstract

Background: When conducing a non-inferiority Phase-III trial, regulatory agencies and investigators might want to get reliable information about rare but serious safety outcomes during the trial. Bayesian non-inferiority approaches have been developed, but commonly utilize historical placebo-controlled data to define the margin, depend on a single final analysis, and no recommendation is provided to define the prespecified decision threshold. In this study, we propose a non-inferiority Bayesian approach for sequential monitoring of rare dichotomous safety events incorporating experts' opinions on margins.

Methods: A Bayesian decision criterion was constructed to monitor four safety events during a non-inferiority trial conducted on pregnant women at risk for premature delivery. Based on experts' elicitation, margins were built using mixtures of beta distributions that preserve experts' variability. Non-informative and informative prior distributions and several decision thresholds were evaluated through an extensive sensitivity analysis. The parameters were selected in order to maintain two rates of misclassifications under prespecified rates, that is, trials that wrongly concluded an unacceptable excess in the experimental arm, or otherwise.

Results: The opinions of 44 experts were elicited about each event non-inferiority margins and its relative severity. In the illustrative trial, the maximal misclassification rates were adapted to events' severity. Using those maximal rates, several priors gave good results and one of them was retained for all events. Each event was associated with a specific decision threshold choice, allowing for the consideration of some differences in their prevalence, margins and severity. Our decision rule has been applied to a simulated dataset.

Conclusions: In settings where evidence is lacking and where some rare but serious safety events have to be monitored during non-inferiority trials, we propose a methodology that avoids an arbitrary margin choice and helps in the decision making at each interim analysis. This decision rule is parametrized to consider the rarity and the relative severity of the events and requires a strong collaboration between physicians and the trial statisticians for the benefit of all. This Bayesian approach could be applied as a complement to the frequentist analysis, so both Data Safety Monitoring Boards and investigators can benefit from such an approach.

Keywords: Clinical trial, Non-inferiority, Bayesian inference, Mixture model, Children, Elicitation

*Correspondence: camille.aupiais@inserm.fr
†Moreno Ursino and Sarah Zohar contributed equally to this work.
[1]Inserm, U1138, Equipe 22, Centre de Recherche des Cordeliers, Sorbonne University, University Paris Descartes, 15 rue de l'École de médecine, 75006 Paris, France
[2]University Paris Diderot, Site Villemin, 10 avenue de Verdun, 75010 Paris, France
Full list of author information is available at the end of the article

© The Author(s). 2019 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Background

Non-inferiority (NI) randomized clinical trials aim to demonstrate whether an experimental treatment is not inferior, below a certain pre-specified margin, to the control treatment [1]. This margin should be formulated according to earlier knowledge and clinical relevance [1, 2]. It has been shown, for instance in paediatrics, that the choice is not well-documented in 63% of studies [3]. However, when there is no reliable placebo-controlled historical data, and when conducting such a trial is not ethical due to changes in practices, margins based solely on clinical judgement could be acceptable, if constructed with rigorous methods, such as systematic analysis of several independent experts' opinions.

When conducing a trial, the analysis of some secondary outcomes, in addition to the primary endpoint, might be challenging, as the sample size was not specifically tuned for that. This issue is of particular importance for safety events, and is even more true when considering rare but critical safety outcomes, which might not occur or only a few can be observed. Consequently, these individual trials are usually under-powered to detect safety differences and to ensure reliable conclusions. Some efficient methods have been proposed for the detection of rare events that are using meta-analysis tools in order to improve overall power. Nevertheless, many methods of meta-analysis are based on large sample approximations, and may be unsuitable when events are rare [4]. Moreover, regulatory agencies and investigators may not wish to wait for post-marketing studies to draw conclusions about rare but serious outcomes of a new intervention. Furthermore, they might want to get reliable safety information before the end of a trial.

When considering NI trials, investigators would like to monitor whether the difference in safety outcomes between arms is clinically relevant. In this case, similar reasoning as for the efficacy primary outcome can be applied, using specific NI margins. If we consider settings where events are rare, a Bayesian approach may seem appropriate to construct sequential stopping rules. Several authors have proposed Bayesian designs for NI trials [5, 6]. Gamalo et al. have proposed a Bayesian NI approach for binary endpoints in which an active-control's treatment effect is estimated using historical data under a fixed margin assumption [7]. However, this Bayesian decision criterion utilizes historical placebo-controlled data, it depends on a single final analysis, and no recommendation is provided to define the prespecified decision threshold.

We propose a Bayesian NI sequential design to monitor several safety dichotomous events where margins are based on clinical relevance obtained from several experts.

Motivation

The ongoing BETADOSE study (NCT02897076) aims to demonstrate that a 50% reduced betamethasone dose regimen is not inferior to a full-dose in preventing neonatal severe respiratory distress syndrome [8]. Several studies have proven the benefice of antenatal corticosteroids, such as betamethasone, so it is used worldwide in pregnant women at risk [9–13]. However, concerns persist regarding long-term adverse events of antenatal corticosteroids, mainly dose-related [14–16].

The trial plans to include 1571 women per arm in 37 French centres. A sequential data analysis has been planned after every 300 newborns reach the primary outcome.

As a safety secondary objective, the protocol plans to monitor, at each interim analysis, the absence of an excess of four other neonatal complications, i.e., neonatal death, severe intraventricular haemorrhage (IVH), necrotising enterocolitis and retinopathy, in two gestational age subgroups of neonates (<28 weeks, 28–32 weeks).

Because only 33% of the randomized women are expected to deliver before 32 weeks, and due to the low frequency of some complications in preterm children, the trial planning had to cope with an expected low number of some secondary events (based on the EPIPAGE-2 cohort study - Additional file 1) [17]. As a consequence, standard frequentist analysis of those outcomes, consisting of tests repeated at each interim analysis, might be powerless.

The Bayesian approach proposed in the manuscript will be applied to this trial (complementary to the frequentist analysis) so the Data Safety Monitoring Board and the investigators can evaluate the difference in terms of the result's interpretation and the benefit of such an approach.

Methods

Let $i = 0, 1$ be the arm-index (1 for the half-dose, 0 for the full-dose) and $j = 1, 2, 3, 4$ the event-index. For the sake of clarity, we show the methodology and results for only one subgroup of neonates (<28 weeks), as it can be repeated in the other subgroup. We used a Bayesian Non-Inferiority approach, detailed in the next subsection. If $\pi_{i,j}$ denotes the event rate in the ith arm, and $D_j \in (0, 1)$ the maximal acceptable difference, the probabilities of interest are $Pr(\pi_{1,j} - \pi_{0,j} \leq D_j)$. To consider the difference in event prevalence and relative severities, this approach was done for each event (j). In our setting, the quantity D_j is not a fixed value, but rather a distribution fitted from elicited experts' opinions through a mixture of beta distributions to consider some variability between experts. The setting of prior distributions and decision thresholds are detailed in the following subsections. Then, a practical example is given using a simulated dataset that mimics the trial. A summary of the general framework is presented in Fig. 1.

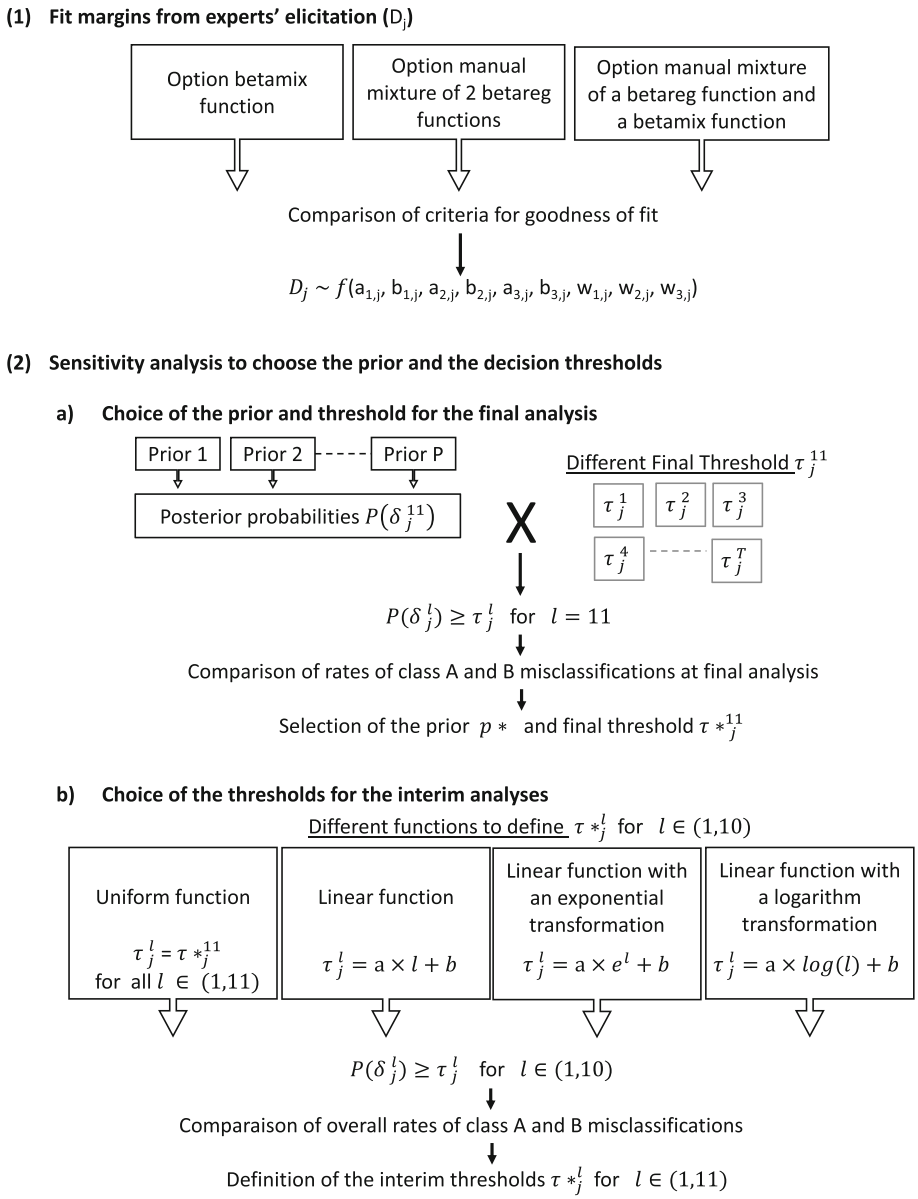

Fig. 1 General framework describing the two steps of the decision rule building. This figure summarizes the general framework, divided in two steps: **1** Fit margins from experts' elicitation (Dj); **2** Sensitivity analysis to choose the prior and the decision thresholds

Bayesian non-inferiority approach

For each event j and arm i, let $y_{i,j,n}$ denote the observed binary outcome for the nth subject, n_i the total number of observations and $Y_{i,j} = \sum_{n=0}^{n_i} y_{i,j,n}$ the number of events. Following a Bayesian binomial model, we have

$$Y_{i,j} \sim Bin(n_i, \theta_{i,j}) \tag{1}$$

where $\theta_{i,j} \sim Beta(\alpha_{i,j}, \beta_{i,j})$ are considered as random variables following a beta prior density. In this setting, the posterior distribution of each $\theta_{i,j}$ is given by:

$$\theta_{i,j}|Y_{i,j} \sim Beta(\alpha_{i,j} + Y_{i,j}, \beta_{i,j} + n_{i,j} - Y_{i,j}) \tag{2}$$

Indexing by l the interim analysis, $l \in [1, \ldots, L]$, we will calculate for each event at each analysis the posterior probability that the difference of events rates, $\theta_{1,j} - \theta_{0,j}$, is higher than the acceptable difference distribution D_j:

Table 1 Three methods of fitting used to model the physicians' acceptable differences of rates of events

1	*Option* `betamix` *function*: For each pair (j,k), application of the `betamix` function with 3 as maximal number of components of the finite mixture.
2	*Option manual mixture of 2* `betareg` *function*: The levels of the observed values of $d_{j,k,e}$ were dichotomized. Then, we fit 2 *Beta* distribution by applying the `betareg` function (or the equivalent `betamix` function with 1 as the number of components) on each level of dichotomisation. All levels of dichotomisation were compared, from that separating the two left values from the others, to that separating the two right values from the others. The two distributions were then mixed by applying the weights w_{1_j} and $w_{2_j} = 1 - w_{1_j}$ to each distribution. The weights $w_{1_j} \in (0, 0.05, 0.10, 0.15, \ldots, 0.95, 1)$ were tested. The models obtained with the different levels of dichotomisation and with the different weights were compared using the criteria for goodness of fit described in Section 1 of the Additional file 7. The fit with the lowest criteria was retained for the comparison with the other 2 methods.
3	*Option manual mixture of a* `betamix` *function and a* `betareg` *function*: A mixture of `betamix` *function* and *manual mixture*: We mixed: (i) a first *Beta* distribution obtained on the left level of dichotomisation (the one obtained with method 2), (ii) a mixture of a second and a third distribution, obtained by applying to the right level of dichotomisation the `betamix` function with 2 as the number of components. The weights given to those distributions were: (i) for the first distribution the w_{1_j} was obtained through method 2, (ii) for the second and third distribution, the weights w_{2_j} and w_{3_j} were obtained through the 'betamix' procedure, multiplied by $(1 - w_{1_j})$.

$$P\left(\delta_j^l\right) = P\left(\theta_{1,j} - \theta_{0,j} > D_j \mid Y_{1,j}^l, Y_{0,j}^l\right) = \int_0^1 \left(\theta_{1,j} - \theta_{0,j} > x \mid Y_{1,j}^l, Y_{0,j}^l, D_j = x\right) . P(D_j = x) . \mathrm{d}x \quad (3)$$

At the lth interim analysis, the Bayesian decision rule will conclude that there is an unacceptable excess of event j in the experimental arm if $P(\delta_j^l) \geq \tau_j^l$, where τ_j^l is a prespecified decision threshold.

Fit margins from experts' elicitation

To evaluate the distribution of D_j, the acceptable difference of events rate between arms, we performed a formal elicitation with several experts. A questionnaire was sent to the two main investigators (1 obstetrician and 1 neonatologist) of each centre involved in the trial. They were asked about (i) their own characteristics (age, sex, speciality, etc.), (ii) the maximum prevalence of events they may tolerate in the experimental arm, given the expected prevalence of each event in the control arm, (iii) the weight of each event, that is the relative severity of the outcomes, considering that death has maximum weight equal to 100.

Let $\tilde{f}_j$ denote the estimated event rate in the full-dose arm, based on the EPIPAGE-2 study (Additional file 1), and $h_{j,e}$ the acceptable event rate in the half-dose arm according to the eth expert, $e \in [1, \ldots, E]$. The acceptable difference between arms according to the eth expert is: $d_{j,e} = h_{j,e} - \tilde{f}_j$. For each event, the distribution of the acceptable difference among the E experts was modeled using a mixture of beta distributions, with a maximum of 3 distributions. Using the `betamix` function (`betareg` package on `R` software [18, 19]), 3 different estimation methods were adopted (the first mathematically driven and the other two empirically driven). See Table 1 for details as well as Section 1 of the Additional file 7. As results, the distribution of D_j will be denoted as $D_j \sim f(a_{1,j}, b_{1,j}, a_{2,j}, b_{2,j}, a_{3,j}, b_{3,j}, w_{1,j}, w_{2,j}, w_{3,j})$, where $(a_{1,j}, b_{1,j})$, $(a_{2,j}, b_{2,j})$ and $(a_{3,j}, b_{3,j})$ are parameters for the 3 *beta* distributions, and $(w_{1,j}, w_{2,j}, w_{3,j})$ the corresponding weights. Parameters will be omitted when mixtures contain less than 3 distributions.

Sensitivity analysis to select the prior and the decision thresholds

The sensitivity analysis aimed to compare the performances of different priors and thresholds τ_j^l and to select the most appropriate combination. In the reference arm, $\theta_{0,j}$ was imputed from historical data (Table 2) [17]. For the experimental arm, five scenarios were considered, determined by the assumed true values of the response probabilities $(\theta_{1,j})$. Let s be the scenario-index $(s \in [1, \ldots, 5])$, and $\theta_{1,j,s}$ denote the prevalence in the experimental arm of the sth scenario. In the first scenario, the prevalence in the experimental and control arms are the same $(\theta_{1,j,1} = \theta_{0,j})$. In the second scenario, the prevalence are lower in the experimental than in the control arm $(\theta_{1,j,2} = 2/3 \times \theta_{0,j})$. In the third to fifth scenario, the prevalence is higher in the experimental than in the control arm $(\theta_{1,j,3} = 1.5 \times \theta_{0,j}$, $\theta_{1,j,4} = 2 \times \theta_{0,j}$ and $\theta_{1,j,5} = 3 \times \theta_{0,j})$. For each scenario, 1000 trials have been generated, with $n_i = 162$ (Additional file 1), and $Y_{i,j,s}$ following the Eq. (1).

The observations of each trial were sampled in L interim analyses. At each analysis, the analysis' population will include the patients of the actual analysis and the patients of the $l - 1$ previous analyses.

To address the issues of how prior location and precision may affect posterior inferences, we constructed an array of P alternative priors, each obtained by specifying numerical values of two quantities, one that changes the prior's location $E(\pi_{1,j} - \pi_{0,j})$ and one that changes its precision (see more details in Section 2 of the Additional file 7).

Choice of the prior and thresholds for the final analysis

The posterior distributions of $\theta_{1,j} - \theta_{0,j}$ of the final L^{th} analysis, were obtained through the Hamiltonian-Monte Carlo method, using the `rstan` package [20, 21] carried out in `R` among the 5000 simulated trials. The posterior probability that it is higher than the acceptable difference distribution was calculated.

Table 2 Prevalence of events assumed in each trial, according to the scenario and to the application data set, and weight and maximal rates of misclassifications assigned to each event to build the decision rule

		Event			
		Death	IVH [1]	NEC [2]	Retinopathy
Simulation study					
All scenarios	Prevalence in FD arm	0.39	0.15	0.06	0.04
Scenario A	Prevalence in HD arm	0.39	0.15	0.06	0.04
	Good decision [3]	Acc	Acc	Acc	Acc
Scenario B	Prevalence in HD arm	0.26	0.10	0.04	0.03
	Good decision [3]	Acc	Acc	Acc	Acc
Scenario C	Prevalence in HD arm	0.58	0.23	0.09	0.06
	Good decision [3]	U	U	Acc	Acc
Scenario D	Prevalence in HD arm	0.78	0.30	0.12	0.08
	Good decision [3]	U	U	U	U
Scenario E	Prevalence in HD arm	1.00	0.45	0.18	0.12
	Good decision [3]	U	U	U	U
Weight[4]		100	88	70	60
Maximal misclassifications rates					
Class *a* misclassifications [5]		0.10	0.10	0.10	0.10
Class *b* misclassifications [6]		0.10	0.16	0.25	0.30
Data set for application					
Prevalence in FD arm		0.39	0.15	0.06	0.04
Prevalence in HD arm		0.47	0.23	0.12	0.08
Good decision [3]		U	U	U	U

HD arm: half-dose arm; FD arm: full-dose arm
[1]IVH: Intraventricular haemorrhage
[2]NEC: Necrotizing enterocolitis
[3]Good decision= What have been considered as good decision for each scenario and event: "Acc" if the difference of prevalence of events is *Acceptable*, "U" if the difference is *Unacceptable*
[4]Weight = Relative severity of the event according to the experts
[5]Class *a* misclassifications rate: Trials that conclude that the difference between arms is *Unacceptable*, among trials with acceptable difference
[6]Class *b* misclassifications rate: Trials that conclude that the difference between arms is *Acceptable*, among trials with unacceptable difference

Then, we calculated, the overall number of misclassifications obtained when applying the decision rule with different decision thresholds τ_j^L at the final analysis, with $\tau_j^L \in (0.50, 1.00)$. Considering the contingency table presented below, we defined two types of misclassifications:

The rates of class *a* and class *b* misclassifications are $= A/(A+C)$ and $= B/(B+D)$, respectively.

This work was repeated for each event, using the *P* priors. Then, the most appropriate prior was selected, along with the thresholds τ_j^L for each event, that is those that

		Truth	
		The difference is Acceptable	The difference is Unacceptable
Conclusion of the decision rule	The difference is Unacceptable	A = Class *a* misclassification	D
	The difference is Acceptable	C	B = Class *b* misclassification

gave acceptable rates of class a and b misclassifications. Let $p*$ denote the selected prior and $\tau*_j^L$ the selected decision thresholds at the L analysis for the event j.

Choice of the thresholds for the interim analyses

To construct the decision rule to be applied at each previous interim analysis, the simulation has been repeated for the L interim analyses, using the $p*$ elected prior. The decision thresholds τ_j^l were defined as follows: (i) for the final analysis, $\tau*_j^L$ was the one defined through the previous step, (ii) for the first analysis, $\tau*_j^1$ has been set to 0.95, (iii) for $l \in (2, L-1)$, four decreasing functions have been tested to define τ_j^l (see Table 3). The overall number of misclassifications obtained with those different functions has been compared. Then, the most appropriate function and thresholds $\tau*_j^l$ have been selected.

Results

Fit margins from experts' elicitation

Among the 78 experts to which the questionnaire was sent, 44 answered (56.4%) (Table 4), including 43 who provided answers about acceptable rates of events in the half-dose arm.

Figure 2 presents the histogram of the acceptable differences of IVH among the E experts ($d_{j,e}$), the fits (D_j) obtained through the 3 different methods, and their criteria for goodness of fit. For the other events, see Additional file 2. The Additional file 3 summarizes the mixtures retained for the acceptable differences D_j.

Sensitivity analysis to select the prior and the decision thresholds

A good sequential decision rule is supposed to help in making a good decision, that is to advise when to stop the trial when the prevalence of events is truly unacceptable and to not stop when the difference is acceptable. Table 2 summarizes what was considered as a "good decision" according to each scenario and events (see more details in the Section 3 of the Additional file 7).

The maximum rate of class a misclassifications has been set to 0.10. For class b misclassifications, we set a maximum inversely proportional to the weight of the event according to the experts (Table 2). Denote by W_j the median weight of the j event among the E experts ($W_j \in [0, 100]$ and $W_{death} = 100$), the maximal rate of class b misclassifications has been set to: $\text{Max(class } b \text{ misclassification)}_j = 0.1 + 0.50 \times \frac{100-W_j}{100}$.

Table 3 Four functions applied to define the thresholds at each of the interim analyses

1 A uniform function: $\tau_j^l = \tau*_j^{11}$ for all $l \in (1, 11)$.

2 A linear function: $\tau_j^l = a \times l + b$,

3 A linear function with an exponential transformation: $\tau_j^l = a \times exp^l + b$,

4 A linear function with a logarithm transformation: $\tau_j^l = a \times log(l) + b$.

Table 4 Main characteristics of the experts who answered to the elicitation questionnaire

Characteristics	N = 44
Age, median (IQR)	46 (39.75-55)
Male sex, n(%)	28 (64)
Number year of being MD [1], median (IQR)	17 (0.75-22.5)
Specialty, n(%)	
Neonatologist	22 (50)
Obstetrician	22 (50)
Type of establishment, n(%)	
University hospital	33 (75)
Position, n(%)	
Hospital practitioner	21 (48)
Professor	21 (48)
Others	2 (4)
History of school of statistics/epidemiology, n(%)	28 (64)
History of being PI [2] of a trial, n(%)	29 (66)

IQR = Interquartile Range
[1] MD: medical doctor
[2] PI: Principal investigator

Selection of the prior and thresholds for the final analysis

Figure 3 shows the number of posterior misclassifications at the final analysis according to each prior and final threshold for IVH. See Additional file 4 for the other events. In an effort to construct a homogeneous decision rule, we selected the same prior for all of the events. Several priors gave acceptable rates of misclassifications (prior 1, 3, 4, 5, 8, 9 and 13). We arbitrarily chose the prior Number 9. Conversely, we applied different final thresholds $\tau*_j^L$ for each event, as they are influenced by the prevalence of events and by the acceptable difference δ_j^l (Table 5).

Selection of the thresholds for the interim analyses

In our case-study, we set $L = 11$. The number of misclassifications obtained by applying the 4 functions defining τ_j^l are presented in the Additional file 5. We retained the linear function with an exponential transformation because it maintained the overall rate of misclassifications under the prespecified acceptable rates. The 3 other functions increased the rate of class a misclassifications over 0.10.

Table 5 summarizes the thresholds finally retained in the decision rule, $\tau*_j^l$, and the overall rates of misclassifications. Figure 4 gives the distribution of the conclusions and misclassifications among the trials, at each interim analysis and in total, for IVH. Additional file 6 represents the distribution of the conclusions and misclassifications for the other events. Finally, Fig. 5 presents the overall numbers or misclassifications obtained by applying this decision rule, according to the scenario.

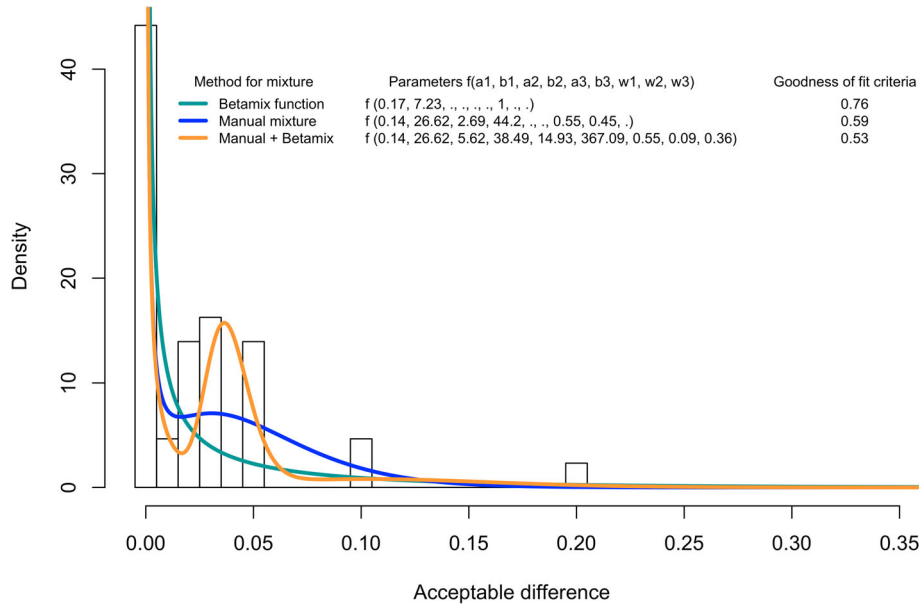

Fig. 2 Histogram of the acceptable difference of severe intraventricular haemorrhage between arms, and mixtures of Beta distributions fitted from experts' elicitation, through 3 different methods, with their criteria for goodness of fit. The histogram represents the acceptable difference of IVH among the E experts ($d_{j,e}$). The 3 lines represent the fits of this difference (D_j), obtained through the 3 different methods. The legend gives the parameters of the fits and their criteria for goodness of fit

Application to data

We applied our method to a simulated dataset for the BETADOSE trial. In this dataset, we considered that the final sample size was 1571 per arm, with $n_i = 162$ for children born before 28 weeks. The prevalence of events was sampled as detailed in Table 2. For all events, a good decision of this trial was considered to conclude an "Unacceptable" difference using the same explanation given before (Section 3 of the Additional file 7). Table 6 summarizes the results at the end of the trial (expressed as observed prevalence) and the Bayesian sequential results, using the rule built in the previous step (Table 5).

At the 6th analysis, since the posterior probabilities became higher than the prespecified threshold $\tau*_j^6$ for death, the trial was stopped because of a potential unacceptable increase of deaths in the experimental arm. If the trial had continued, it would have stopped at the 10th analysis because of IVH.

Discussion

Motivated by the desire to deal with settings where rare but serious events have to be monitored during an non-inferiority trial, we have proposed a methodology that provides a practical way to help in the decision making at each interim analysis.

Our approach has the advantage of incorporating experts' opinions about the non-inferiority margins. As a consequence, it can be used as an alternative in cases where historical placebo-controlled data aren't available. We have proposed to keep the variability among experts and used a distribution instead of a discrete margin. Indeed, we could have averaged all experts' opinions, but this will not have reflected all potential variability. In a simulation study, we compared our approach to the use of average values (see Additional file 9). We found that the use of a mixture gave different results than the use of the mean of the experts' opinions. Indeed, the difference between the two approaches increased as the variability among experts increased. Moreover, we could have weighted experts' opinions according to some pertinent covariates. In a previous work, Thall et al. compared different ways to weight physicians' opinion using mixture priors of the parameter of interest [22]. The authors found, according to their design, that posterior quantities appear to be insensitive to how physicians are weighted, so we decided to weight all physicians equally. In our case, the variability among experts was kept in order to reflect all potential opinions, that is the distribution across all the range of potential margins. Our method can be applied whatever the values are in between zero and one.

One limitation of our motivating example is that the majority of the experts set the acceptable difference to zero, whereas zero is not a possible value for a non-inferiority margin. When generalizing this method to another non-inferiority trial, we suggest to investigators to remind the experts that the margin cannot be set to zero.

Because the prior chosen for a Bayesian analysis needs to be well documented and robust to its parameter

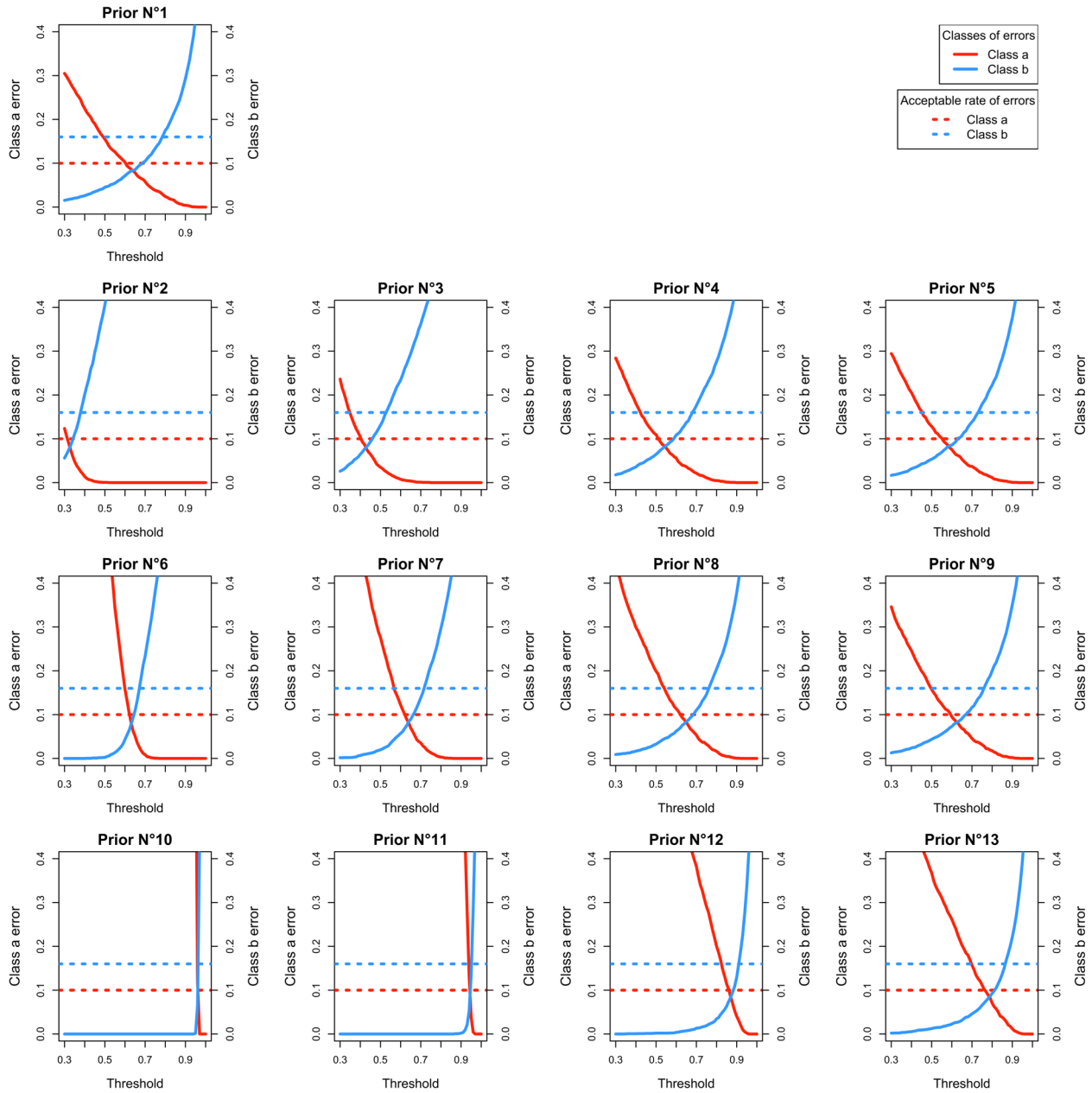

Fig. 3 Plots of posterior class *a* and class *b* misclassifications according to the decision thresholds for each of the 13 pairs of priors for severe intraventricular haemorrhage. This figure represents the posterior rates of misclassifications for each pair of priors. Prior 1 is the non-informative prior, with $\alpha_{1j} = \alpha_{0j} = \beta_{1j} = \beta_{0j} = 1$; Prior 2 to 13 are distinguished by (i) the means for the difference between the two arms: $E(\pi_{1j} - \pi_{0j}) = 0$ for prior 2, 3, 4 and 5; $E(\pi_{1j} - \pi_{0j}) = \text{median}(d_{j,e})$ for prior 6, 7, 8 and 9; and $E(\pi_{1j} - \pi_{0j}) = \pi_{0j}$ for prior 10, 11, 12 and 13; (ii) their precision: 1 for prior 2, 6 and 10; 1/3 for prior 3, 7 and 11; 1/10 for prior 4, 8 and 12; and 1/20 for prior 5, 9 and 13. For each prior, the red solid line represents the number of posterior class *a* misclassifications (trials that conclude that the difference between arms is *Unacceptable*, while it is not true) at the final analysis, according to each final threshold. The blue solid line represents the number of posterior class *b* misclassifications (trials that conclude that the difference between arms is *Acceptable*, while it is not true)

choices, we performed an extensive sensitivity analysis evaluating non-informative and informative priors and several thresholds. Thresholds retained were varying between events, allowing us to consider the differences in prevalence, and in margins and severity conferred by clinicians to each event. Likewise, when we repeated this work in the subgroup of premature infants born after 28 weeks (results not shown), the thresholds were different, reflecting the higher rarity of events and the different margins.

Table 5 Final decision rule retained through the sensitivity analysis: thresholds to be applied at each interim analysis and final overall rates of misclassifications, according to the event

	Event			
	Death	IVH [2]	NEC [3]	Retinopathy
Thresholds $\tau*_j^l$ [1], to applied in analysis				
1	0.95	0.95	0.95	0.95
2	0.95	0.95	0.95	0.95
3	0.95	0.95	0.95	0.95
4	0.95	0.95	0.95	0.95
5	0.949	0.949	0.950	0.949
6	0.948	0.948	0.949	0.948
7	0.944	0.944	0.948	0.946
8	0.934	0.933	0.944	0.938
9	0.907	0.903	0.932	0.918
10	0.832	0.821	0.902	0.862
11	0.63	0.60	0.82	0.71
Overall rate of errors				
Class *a* misclassifications [4]	0.09	0.09	0.10	0.10
Class *b* misclassifications [5]	<0.01	0.07	0.20	0.27

[1] $\tau*_j^l$: threshold to apply in the Bayesian decision rule for the event *k* in the subgroup *j*, at the *l* interim analysis: the rule will conclude that their is an unacceptable excess if $P(\delta_j^l) \geq \tau_j^l$

[2] IVH: Intraventricular haemorrhage

[3] NEC: Necrotizing enterocolitis

[4] Class *a* misclassifications: Trials that conclude that the difference between arms is *Unacceptable*, while it is not true

[5] Class *b* misclassifications: Trials that conclude that the difference between arms is *Acceptable*, while it is not true

To choose the best priors and stopping thresholds, the rates of misclassifications have been computed and compared. As the two types of misclassifications are moving in opposite directions, we had to find a compromise between the two. Since we do not want to wrongly conclude too often an inferiority of the experimental arm, we decided to set a maximum for class A misclassification at 0.10, to be more permissive in terms of class B misclassifications and to adapt this permissiveness to the severity of each event. To define the stopping thresholds at each interim analysis, simulations have compared several initial thresholds and four decreasing functions of τ. The purpose of this study was to find the best thresholds in order to have good functional properties of the design, i.e. do not stop frequently at the beginning when it is wrong and do not continue until the end when we have to stop. Finally, as we dealt with some rare events, overall rates of class A and B misclassifications were relatively high. This has to be put in balance with frequentist type I and type II error rates that sometimes have to be compromised, especially in the case of rare secondary events.

When generalizing this method to another trial, this work needs to be repeated before the analysis of the real data; the maximal rates of class A and B misclassifications

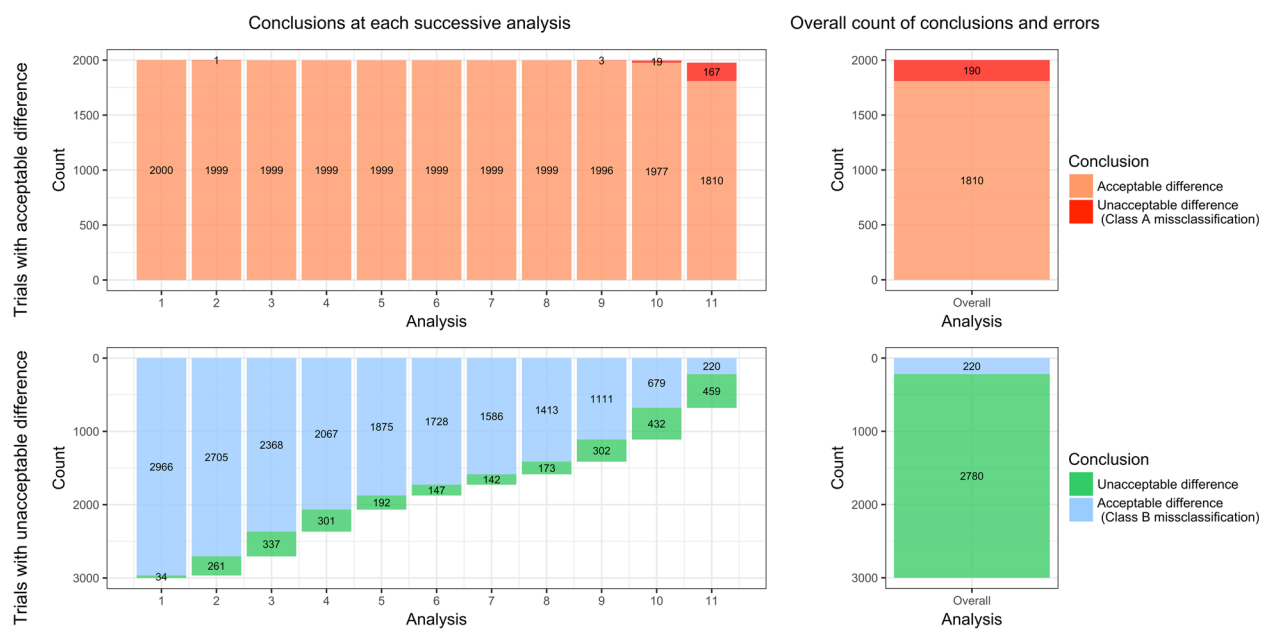

Fig. 4 Distribution of the successive conclusions and errors, obtained by applying the decision rule to the 5000 simulated trials, at each interim analysis and in overall, for severe intraventricular haemorrhage. The left part of the plot represents the conclusions at each interim analysis. The right part represents the overall count of conclusions among the 11 analyses. The upper part of the plot represents the trials with an *Acceptable* difference between arms: orange area correspond to trials that conclude that the difference between arms is *Acceptable*, while it is true; red area correspond to trials that conclude that the difference between arms is *Unacceptable*, while it is not true (class *a* misclassifications). The bottom part of the plot represents the trials with an *Unacceptable* difference between arms: green area correspond to trials that conclude that the difference between arms is *Unacceptable*, while it is true; blue area correspond to trials that conclude that the difference between arms is *Acceptable*, while it is not true (class *b* misclassifications)

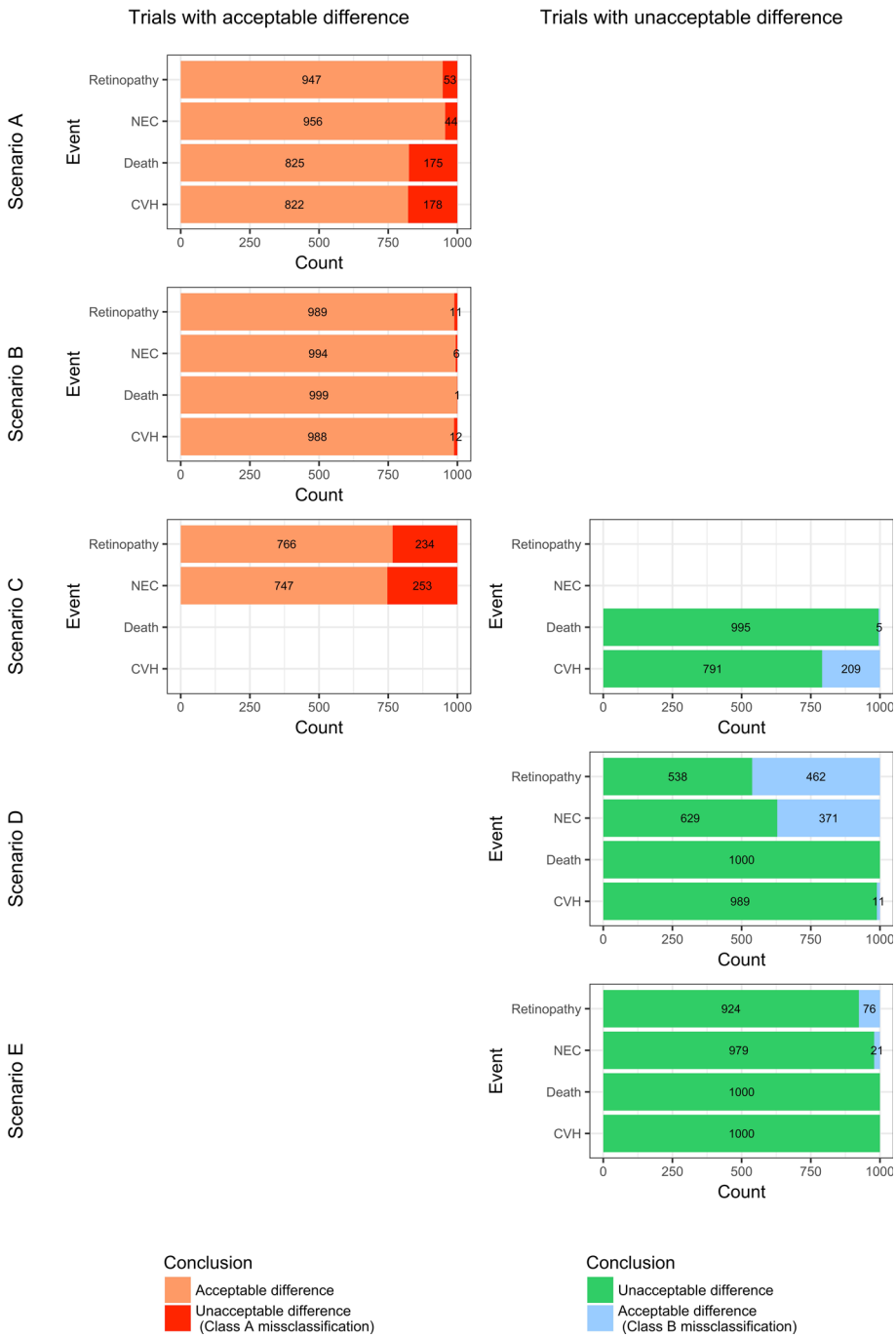

Fig. 5 Distribution of the overall conclusions and errors, obtained by applying the decision rule to the 5000 simulated trials, according to the event and the scenario. This plot presents the overall numbers or misclassifications obtained by applying this decision rule, according to the 5 scenario and to the 4 events. The left part of the plot represents the trials with an *Acceptable* difference between arms: orange area correspond to trials that conclude that the difference between arms is *Acceptable*, while it is true; red area correspond to trials that conclude that the difference between arms is *Unacceptable*, while it is not true (class *a* misclassifications). The right part of the plot represents the trials with an *Unacceptable* difference between arms: green area correspond to trials that conclude that the difference between arms is *Unacceptable*, while it is true; blue area correspond to trials that conclude that the difference between arms is *Acceptable*, while it is not true (class *b* misclassifications). IVH: Intraventricular haemorrhage; NEC: Necrotizing enterocolitis

Table 6 Observations and decision obtained by application of the Bayesian decision rule to a data set

Analysis	N		Death						Intraventricular haemorrhage						Necrotising enterocolitis						Retinopathy					
			Observations				Bayesian		Observations				Bayesian		Observations				Bayesian		Observations				Bayesian	
	FD arm	HD arm	FD arm		HD arm		Post prob.	Decision	FD arm		HD arm		Post prob.	Decision	FD arm		HD arm		Post prob.	Decision	FD arm		HD arm		Post prob.	Decision
			n	%	n	%			n	%	n	%			n	%	n	%			n	%	n	%		
1	10	20	5	0.50	11	0.55	0.60	0	0	0	1	0.05	0.42	0	0	0	1	0.05	0.59	0	0	0	2	0.10	0.69	0
2	25	35	12	0.48	20	0.57	0.66	0	1	0.04	6	0.17	0.71	0	1	0.04	2	0.06	0.55	0	0	0	4	0.11	0.84	0
3	39	51	16	0.41	29	0.57	0.79	0	3	0.08	9	0.18	0.70	0	3	0.08	7	0.14	0.75	0	0	0	5	0.10	0.87	0
4	54	68	22	0.41	42	0.62	0.89	0	4	0.07	13	0.19	0.80	0	5	0.09	10	0.15	0.74	0	0	0	8	0.12	0.95	0
5	77	80	32	0.42	53	0.66	0.94	0	8	0.10	17	0.21	0.81	0	5	0.06	10	0.12	0.79	0	4	0.05	9	0.11	0.78	0
6	88	92	34	0.39	58	0.63	0.95	1	9	0.10	22	0.24	0.89	0	6	0.07	12	0.13	0.81	0	5	0.06	10	0.11	0.76	0
7	99	104	39	0.39	65	0.62	0.95		14	0.14	25	0.24	0.81	0	10	0.10	16	0.15	0.76	0	5	0.05	12	0.12	0.83	0
8	116	121	45	0.39	72	0.60	0.94		16	0.14	31	0.26	0.88	0	10	0.09	18	0.15	0.83	0	6	0.05	12	0.10	0.76	0
9	127	139	50	0.39	80	0.58	0.93		18	0.14	33	0.24	0.84	0	11	0.09	20	0.14	0.83	0	6	0.05	13	0.09	0.76	0
10	143	157	54	0.38	86	0.55	0.93		21	0.15	37	0.24	0.83	1	12	0.08	24	0.15	0.88	0	6	0.04	15	0.10	0.82	0
11	154	163	60	0.39	90	0.55	0.92		24	0.16	37	0.23	0.76		12	0.08	24	0.15	0.89	1	7	0.05	15	0.09	0.78	1

[0]HD arm: half-dose arm; FD arm: full-dose arm

have to be balanced, considering the setting, and the parameters of the decision rule have to be adapted in consequence, namely the prior, the margins and the decision thresholds. Finally, after having pre-specified all these parameters, the decision rule can be applied by the statistician to the unblinded data, and presented to the Data Safety Monitoring Board. In order to apply this methodology, we already designed a non-inferiority trial that should start in few months, using the same statistical approach in an other setting.

In conclusion, our approach was found to be efficient in dealing with safety monitoring of rare and non-rare events in a non-inferiority context. It requires a strong collaboration between physicians and the trial statisticians for the benefit of all.

Conclusion

We proposed a practical way to help to assist with decisions on safety dichotomous events at each interim analysis of a non-inferiority trial. This Bayesian design is suitable for rare events and for non-rare events. It incorporates experts' opinions on margins, so it can be constructed without historical placebo-controlled data. This Bayesian sequential approach could be applied as a complement to the frequentist analysis, so both Data Safety Monitoring Boards and investigators can benefit from such an approach.

Additional files

Additional file 1: Expected distribution of gestational age and events in the control arm (FD arm) of the BETADOSE trial, imputed from the prevalence observed in the ePIPAGE-2 cohort study. A table providing the expected event rates and gestational age in the motivated trial (based on the EPIPAGE-2 cohort study [17]). (PDF 78 kb)

Additional file 2: Histogram of the acceptable differences in events, and mixtures of beta distributions fitted from experts' elicitation, through 3 different methods, with their criteria for goodness of fit. The plots analogous to Fig. 2 for the 3 other events: (a) Death, (b) Necrotizing enterocolitis, (c) Retinopathy. (PDF 992 kb)

Additional file 3: Parameters of the mixtures retained to fit the acceptable differences between arms D_j. A table summarizing the mixtures retained for the acceptable differences. (PDF 108 kb)

Additional file 4: Plots of posterior class *a* and class *b* misclassifications according to the stopping thresholds for each of the 13 pairs of priors. The plots analogous to Fig. 3 for the 3 other events: (a) Death, (b) Necrotizing enterocolitis, (c) Retinopathy. (PDF 2969 kb)

Additional file 5: Final overall rates of misclassifications obtained with the decision rule, according to the function used to define the thresholds at each successive interim analysis. The final overall rates of misclassifications obtained with the decision rule summarized in the Table 5, provided for the 4 functions used to define the thresholds at each successive interim analysis. (PDF 107 kb)

Additional file 6: Distribution of the successive conclusions and misclassifications, obtained by applying the decision rule to the 5000 simulated trials at each interim analysis and in overall. The plots analogous to Fig. 4 for the 3 other events: (a) Death, (b) Necrotizing enterocolitis, (c) Retinopathy. (PDF 2498 kb)

Additional file 7: Supplemental information's on methods. Additional information on methods : (i) Method for comparison of the 3 different estimation ways used to fit the mixtures of Beta distributions; (ii) Definition of the non-informative and informative priors compared in the sensitivity analysis; (iii) Definition of the "good decision" for each scenario of the simulation study. (PDF 246 kb)

Additional file 8: Physician experts who participated in the elicitation, BETADOSE trial. Complete list of experts who participated in the elicitation. (PDF 41 kb)

Additional file 9: Comparison of the use of a mixture distribution versus the use of an average value of elicitated margins from experts. In this additional work, two approaches were proposed to compute the acceptable difference from experts' elicitation: the method described in the main manuscript and the use of average values. (PDF 368 kb)

Additional file 10: Example of the code to simulate the results for one event (neonatal death): In the part A, it fit margins from experts' elicitation, using a fictive data set of experts' answers (E1). In the part B, it simulate the trials and compute the differences of the posterior samplers for M pairs, for one scenario, using one informative prior. In the part C, it calculate the posterior probability that the difference of rate of events is higher than the acceptable difference according to experts, and compute the decision using a threshold of 0.50. (R 26 kb)

Abbreviations

IVH: Intraventricular haemorrhage; NI: Non-inferiority

Acknowledgements

The authors thank Bruno Giraudeau and Rym Boulkedid for assistance in preparing the elicitation questionnaire. They also thank all the physician experts who participated in the elicitation (see the complete list in Additional file 8).

Authors' contributions

CAu conceived the concept of this study, carried out the simulations and the analyses, interpreted the results, and drafted the manuscript. CAl designed the motivated trial, supervised the work, critically reviewed and made substantial contributions to the manuscript. TS designed the motivated trial, critically reviewed and made substantial contributions to the manuscript. OB designed the motivated trial, critically reviewed and made substantial contributions to the manuscript. MU conceived the concept of this study, critically reviewed and made substantial contributions to the manuscript. SZ conceived the concept of this study, critically reviewed and made substantial contributions to the manuscript. All authors have seen, commented on and approved the final manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-f or-profit sectors. The BETADOSE study is supported by a research grant from the French Health Ministry and sponsored by the Direction de la Recherche et de l'Innovation, APHP (Programme Hospitalier de Recherche Clinique, AOM15158) after a peer review process, but the research work covered in this manuscript has been conducted independently from this funding of the BETADOSE trial.

Availability of data and materials

We provided as a supplementary material an example of the code to run the 2 steps of this work (Additional file 10).The complete materials used for this study (R code and dataset generated) are available from the corresponding author on reasonable request.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declared declare that they have no competing interests with respect to the research, authorship, and/or publication of this article.

Author details
[1]Inserm, U1138, Equipe 22, Centre de Recherche des Cordeliers, Sorbonne University, University Paris Descartes, 15 rue de l'École de médecine, 75006 Paris, France. [2]University Paris Diderot, Site Villemin, 10 avenue de Verdun, 75010 Paris, France. [3]Inserm, U1123, ECEVE, 10 avenue de Verdun, 75010 Paris, France. [4]Unité d'épidémiologie clinique, CIC-EC 1426, Assistance Publique-Hôpitaux de Paris, Hôpital Robert Debré, 48 boulevard Sérurier, 75019 Paris, France. [5]Service de Gynécologie Obstétrique, Assistance Publique-Hôpitaux de Paris, Hôpital Robert Debré, 48 boulevard Sérurier, 75019 Paris, France. [6]Inserm, U1153, Epidemiology and Biostatistics Sorbonne Paris Cité Research Center, Obstetrical, Perinatal and Pediatric Epidemiology Team, 53 avenue de l'observatoire, 75014 Paris, France. [7]Service de néonatalogie, Assistance Publique-Hôpitaux de Paris, Hôpital Robert Debré, 48 boulevard Sérurier, 75019 Paris, France. [8]Inserm, U1141, Hôpital Robert Debré, 48 boulevard Sérurier, 75019 Paris, France. [9](Present address) Service de néonatalogie, Hôpitaux universitaires de Genève, 32 boulevard de la Cluse, 1205 Genève, Suisse. [10]F-CRIN PARTNERS Platform (AP-HP), 10 avenue de Verdun, 75010 Paris, France.

Received: 22 January 2019 Accepted: 21 August 2019
Published online: 18 September 2019

References

1. U S Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER). Non-Inferiority Clinical Trials to Establish Effectiveness - Guidance for Industry; 2016.
2. Sorbello A, Komo S, Valappil T. Noninferiority Margin for Clinical Trials of Antibacterial Drugs for Nosocomial Pneumonia. Drug Inf J. 2010;44(2): 165–76.
3. Aupiais C, Zohar S, Taverny G, Le Roux E, Boulkedid R, Alberti C. Exploring how non-inferiority and equivalence are assessed in paediatrics: a systematic review. Arch Dis Child. 2018;103(11):1067–75.
4. Bradburn MJ, Deeks JJ, Berlin JA, Russell Localio A. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. Stat Med. 2007;26(1):53–77.
5. Gamalo-Siebers M, Gao A, Lakshminarayanan M, Liu G, Natanegara F, Railkar R, et al. Bayesian methods for the design and analysis of noninferiority trials. J Biopharm Stat. 2016;26(5):823–41.
6. Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian Approaches to Randomized Trials. J R Stat Soc Ser A (Stat Soc). 1994;157(3):357–416.
7. Gamalo MA, Wu R, Tiwari RC. Bayesian approach to noninferiority trials for proportions. J Biopharm Stat. 2011;21(5):902–19.
8. Schmitz T, Alberti C, Ursino M, Baud O, Aupiais C, BETADOSE study group and the GROG(GroupedeRechercheenGynécologie Obstétrique). Full versus half dose of antenatal betamethasone to prevent severe neonatal respiratory distress syndrome associated with preterm birth: study protocol for a randomised, multicenter, double blind, placebo-controlled, non-inferiority trial (BETADOSE). BMC Pregnancy Childbirth. 2019;19(1):67.
9. Liggins GC, Howie RN. A controlled trial of antepartum glucocorticoid treatment for prevention of the respiratory distress syndrome in premature infants. Pediatrics. 1972;50(4):515–25.
10. Effect of corticosteroids for fetal maturation on perinatal outcomes. NIH Consens Statement. 1994;12(2):1–24.
11. ACOG committee opinion. Antenatal corticosteroid therapy for fetal maturation. Number 210, October 1998 (Replaces Number 147, December 1994). Committee on Obstetric Practice. American College of Obstetricians and Gynecologists. Int J Gynaecol Obstet Off Organ Int Fed Gynaecol Obstet. 1999;64(3):334–5.
12. Senat MV. Corticosteroid for fetal lung maturation: indication and treatment protocols. J Gynecol Obstet Biol Reprod. 2002;31(7 Suppl): 5S105–113.
13. Royal College of Obstetricians & Gynaecologists. Green-top Guideline No 7: Antenatal corticosteroids to reduce neonatal morbidity and mortality. London; 2010.
14. Roberts D, Dalziel S. Antenatal corticosteroids for accelerating fetal lung maturation for women at risk of preterm birth. Cochrane Database Syst Rev. 2006;3:CD004454. http://dx.doi.org/10.1002/14651858.cd004454.pub2.
15. Brownfoot FC, Gagliardi DI, Bain E, Middleton P, Crowther CA. Different corticosteroids and regimens for accelerating fetal lung maturation for women at risk of preterm birth. Cochrane Database Syst Rev. 2013;8: CD006764. http://dx.doi.org/10.1002/14651858.cd006764.
16. Crowther A, McKinlay CJD, Middleton P, Harding JE. Repeat doses of prenatal corticosteroids for women at risk of preterm birth for improving neonatal health outcomes. Cochrane Database Syst Rev. 2015;7: CD003935. http://dx.doi.org/10.1002/14651858.cd003935.pub3.
17. Ancel PY, Goffinet F, EPIPAGE-2 Writing Group, Kuhn P, Langer B, Matis J, et al. Survival and morbidity of preterm children born at 22 through 34 weeks' gestation in France in 2011: results of the EPIPAGE-2 cohort study. JAMA Pediatr. 2015;169(3):230–8.
18. R Core Team. R: A Language and Environment for Statistical Computing; R Foundation for Statistical Computing; 2017. https://www.R-project.org/.
19. Grün B, Kosmidis I, Zeileis A. Extended Beta Regression in R: Shaken, Stirred, Mixed, and Partitioned. J Stat Softw. 2012;48(11):1–25.
20. Stan Development Team. RStan: the R interface to Stan. 2018. R package version 2.11.1. http://mc-stan.org/.
21. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: A Probabilistic Programming Language. J Stat Softw. 2017;76(1):1–32.
22. Thall PF, Ursino, M, Baudouin V, Alberti C, Zohar S. Bayesian treatment comparison using parametric mixture priors computed from elicited histograms. Stat Methods Med Res. 2019;28(2):404–18.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC