BJA

British Journal of Anaesthesia, 126 (2): 550–555 (2021)

doi: 10.1016/j.bja.2020.08.062
Advance Access Publication Date: 29 October 2020
Special Article

STATISTICS

Bayesian predictive probabilities: a good way to monitor clinical trials

David Ferreira[1,2,*], Pierre-Olivier Ludes[3], Pierre Diemunsch[3], Eric Noll[3], Klaus D. Torp[4] and Nicolas Meyer[2,5]

[1]Anesthesiology and Intensive Care Department, Centre Hospitalier Universitaire de Besançon, Besançon, France, [2]iCUBE, UMR7357, Université de Strasbourg, Illkirch Cedex, France, [3]Anesthesiology and Intensive Care Department, IHU-Strasbourg, Centre Hospitalier Universitaire de Strasbourg, Strasbourg, France, [4]Department of Anesthesiology, Mayo Clinic, Jacksonville, FL, USA and [5]Public Health Department, Groupe de Méthodes en Recherche Clinique, Centre Hospitalier Universitaire de Strasbourg, Strasbourg, France

*Corresponding author. E-mail: dferreira@chu-besancon.fr

Abstract

Background: Bayesian methods, with the predictive probability (PredP), allow multiple interim analyses with interim posterior probability (PostP) computation, without the need to correct for multiple looks at the data. The objective of this paper was to illustrate the use of PredP by simulating a sequential analysis of a clinical trial.
Methods: We used data from the Laryngobloc trial that planned to include 480 patients to demonstrate the equivalence of success between a laryngoscopy performed with the Laryngobloc® device and a control device. A crossover Bayesian design was used. The success rates of the two laryngoscopy devices were compared. Interim analyses, computed from random numbers of subjects, were simulated.
Results: The PostP of equivalence rapidly reached the predefined bound of 0.95. The PredP computed with an equivalence margin of 10% reached the efficacy bound between 352 and 409 of the 480 included patients. If a frequentist analysis had been made on the basis of 217 out of 480 subjects, the study would have been prematurely stopped for equivalence. The PredP indicated that this result was nonetheless unstable and that the equivalence was, thus far, not guaranteed.
Conclusions: Based on these interim analyses, we can conclude with a sufficiently high probability that the equivalence would have been met on the primary outcome before the predetermined end of this particular trial. If a Bayesian approach using PredP had been used, it would have allowed an early termination of the trial by reducing the calculated sample size by 15–20%.

Keywords: Bayesian statistics; clinical trial; monitoring; predictive probabilities; RCT; statistics

Editor’s key points

- Bayesian statistical methods equate more closely to how clinicians think and make decisions.
- Unlike frequentist statistics, Bayesian inference updates the probability of an occurrence as more information becomes available.
- Prior beliefs or knowledge is overtly included in a summation that is updated by new data to create ‘posterior’ beliefs or knowledge, reducing uncertainty.
- Bayesian predictive probabilities can schematically describe the "stability" of the data in an interim analysis by considering all possible future data, and thus provide support to authors in prematurely stopping a trial.

Received: 16 September 2019; **Accepted:** 30 August 2020

Statistical analyses for clinical studies are usually conducted using the frequentist or classical methods. Frequentist strategies of monitoring are based on interim P-values, which cause alpha risk inflation, and the interim analyses strategy must be planned very carefully before beginning the study. Moreover, *P*-value boundaries need complex calculations and they must be defined beforehand for each interim analysis, moving us further from any clinical significance.[1] Indeed, a frequentist interim design cannot be modified in the light of the evidence accumulated during the trial, and the interim analyses must be done at the time planned. The use of stochastic curtailment, methods allowing a rigorous intermediate analysis, can be set only at great computational cost and with heavy consequences on ability to interpret the results that rely on the null hypothesis test and its limitations.[2,3] Any additional unplanned analysis is 'forbidden'. This prevents the trial investigators to take into account any unexpected but relevant information that occurs during the course of the trial, such as high efficacy or an important level of adverse effects.[4] Indeed, an unanticipated high toxicity rate that modifies the course of the trial cannot be properly managed in the analysis, whilst a Bayesian inference on such modified trials is valid.[4,5]

Bayesian methods are very flexible tools, allowing a much simpler implementation of sequential analysis.

The use of the principle of Bayesian methods requires to have some knowledge about the interest parameter (e.g. a mean difference).[6,7] This knowledge is more or less precise, but is expressed in the form of a probability distribution, which indicates the probability that the parameter will take on a certain value. This distribution is called the *a priori* distribution or simply 'the prior'. Pathophysiological knowledge of the phenomenon or previous studies often provides a good estimate of this prior distribution. This probability distribution then makes it possible to calculate the probability of observing the data obtained during a clinical trial. This is a calculation close to, but different from, that of the *P*-value, which is traditionally used. The knowledge provided by the data is then combined with the prior knowledge on the parameter to obtain a so-called *a posteriori* or posterior probability distribution, which contains everything we know about the parameter of interest (mean difference or any other parameter) after the study is carried out. This probabilistic knowledge about the parameter is therefore increased simply by a manipulation on probabilities.

Amongst the different Bayesian tools used to monitor a sequential trial, the predictive probability (PredP) and the posterior probability (PostP) are of particular interest. They nevertheless answer different questions. For instance, in an interim analysis of an equivalence design, the PostP of equivalence is the probability that the two devices are equivalent, based on the data available at the time of this interim analysis. In contrast, PredP is the probability that the two devices are considered equivalent on the future, not yet observed, final observations, as computed in the sample size, conditional on the observed data at the time of the interim analysis. In other words, PredP is the probability of observing a specific future outcome (not necessarily limited to the current conclusion) based on the current knowledge (as summarised in the current posterior probability distribution). Interim analysis then opens the way to early trial stopping for futility or efficacy. Futility means there is little chance that the study reaches a predefined effect size with a high probability, and one can consider to stop the study. Efficacy means that the study is highly likely to reach a predefined effect size with a high probability, suggesting the inclusions could be stopped.

As for any Bayesian procedure, PostP and PredP can be computed at any time in the course of the trial, even if the times have not been pre-specified in the protocol. Thus, the Bayesian method allows performance of multiple interim analyses with recurrent interim PostP computation, without the need to correct for multiple looks at the data.[8] Given the multiple advantages, the US Food and Drug Administration has issued a guidance for using Bayesian methods in medical device clinical trials.[9] Despite this guidance, Bayesian methods are still rarely used in current medical research for Phase III trials, probably because of the lack of physician knowledge and training.[10] However, its use is increasingly common for early-phase trials in drug development, particularly in oncology.

The aim of this study was to illustrate the use of PredP during a clinical trial comparing the equivalence of two laryngoscopes. Our hypothesis was that the Bayesian method would provide better guidance to the timing of study termination.

Methods

To illustrate our comments, we used the Laryngobloc study that is currently being submitted for publication (clinical trial registration ID NCT01632085). The main objective of this RCT was to demonstrate the equivalence of success between a laryngoscopy performed with the Laryngobloc® device (LB; VBM Paris, France) and a control device. As the order of use of the laryngoscopic devices may influence the evaluation of the primary outcome because of uncontrollable procedure-specific criteria (i.e. the assumption that the second laryngoscopy may be easier than the first), a two-period two-sequence crossover design was used. The success rates of the two laryngoscopy devices were compared. A mixed model showed that there was neither period nor order effect. We thus combined the results of the two periods in a single table (Table 1).

Two groups were randomised as follows:

(i) R group: a first laryngoscopy was performed with the control device (i.e. the single-use Macintosh metal blade (SG Manufacturers, Sialkot, Pakistan) and the reusable handle). The second laryngoscopy was performed with the LB, a single-use laryngoscope of the same Macintosh blade design. This device was named after its single block

Table 1 Description of the possible distribution of study results. CL, Cormack and Lehane grade; LB device, Laryngobloc® device; *P*, probability; R device, control device; *X*, *n*, number of patients.

Device	LB		Total
	Success (1) CL 1–2	**Failure (0) CL 3–4**	
R Success (1) CL 1–2	X_{11} (P_{11})	X_{10} (P_{10})	$X_{1.}$ ($P_{1.}$)
Failure (0) CL 3–4	X_{01} (P_{01})	X_{00} (P_{00})	$n-X_{0.}$ ($P_{0.}$)
Total	$X_{.1}$ ($P_{.1}$)	$n-X_{.0}$ ($P_{.0}$)	N

structure with the blade and the handle consisting of plastic and forming one and the same part, with no hinge, avoiding folding the blade over the handle.

(ii) LB group: a first laryngoscopy was performed with the LB and the second one was performed with the control device. A Cormack and Lehane classification Grade 1 or 2 was considered as a success and a Grade 3 or 4 was considered as a failure.

Statistical analysis

The hypothesis of the trial was that the laryngoscopy with the LB and the control device were equivalent in terms of glottic visualisation success rates with a PostP of equivalence above a threshold of 0.95. This very high threshold was chosen because of the need to have a high level of confidence in the conclusion for this procedure, which is extremely common in daily clinical practice, in a study with low risk for the patient. In this paper, for illustrative purpose, we arbitrarily chose a range of equivalence of plus or minus 10% on the proportion of laryngoscopy success (Grade 1 or 2). Statistically speaking, this equivalence is expressed as a proportion of discordant outcome that must be less than 10% in absolute value. In Table 1, which displays the parameters for the paired binary outcomes, equivalence occurs when $D=P_{10}+P_{01}<10\%$[11] (Fig. 1). The estimation of the Cormack and Lehane score on the first and second attempts (two successive visualisations, one with each device, on the same subject) was performed.

A sample size determination was computed using a classical method, which does not hamper the use of Bayesian method at the time of analysis.[12] A sample size of 457 patients was required and was increased to 480 to compensate for potential missing data.

Principles and realisation of the statistical analysis for the PredP-based interim analysis

We performed a Bayesian analysis for paired categorical data. In our example, there were four possible results over the two laryngoscopies, as described in the contingency table (Table 1).

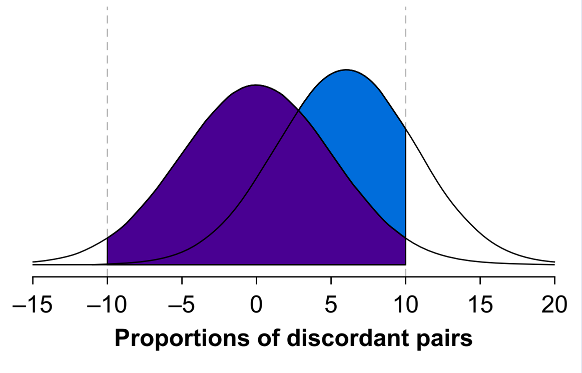

Fig 1. Description of the posterior probability (PostP) of equivalence to illustrate the concepts of the 'PostP equivalence threshold' (0.95) and the 'equivalence range' ($D=P_{10}+P_{01}$ <10%). The curve centred on 0 fall within the 10% range almost completely (95% of the curve) (non-equivalence rejected), whilst the other would have only 80% of the area under the curve within the 10% range (non-equivalence not rejected).

In a Bayesian analysis, parameters must be described with probability distributions (Gaussian distribution for normal data, for instance). The Dirichlet distribution is one of the distributions used in the Bayesian analysis of contingency tables to estimate the parameters (i.e. the probability of being in a given table cell). This distribution has several parameters, which, in the case of a contingency table, are the frequencies in each cell of the table. For a table with frequencies x_{11}, x_{10}, x_{01}, and x_{00}, the distribution is thus a Dirichlet Di(x_{11}; x_{10}; x_{01}; x_{00}). Moreover, this distribution can be used as prior and as posterior distributions, thanks to the properties mentioned previously that the table frequencies are to be interpreted as the distribution parameters. The prior distribution was updated by the interim data to generate an interim PostP distribution. In the case of a contingency table, the posterior distribution is very easily derived by adding the observed frequencies to the prior frequencies. In other words, the posterior distribution for the parameters is easily calculated as the Dirichlet distribution with the hyperparameters equal to the prior 'count' plus the observed count.

The prior parameter distribution was expressed as a table containing a 'number' of subjects of 0.5 in each cell (Dirichlet distribution Di[0.5, 0.5, 0.5, 0.5]). This is equivalent to information from 2(=0.5+0.5+0.5+0.5) patients. This is a way to give a very small weight to the prior so as not to influence heavily the posterior distribution.

A sensitivity analyses was performed here by varying the prior distribution Di(x_{11}; x_{10}; x_{01}; x_{00}) using either Di(1, 1, 1, 1), as a minimally informative prior, or Di(10, 1, 1, 10), favouring slightly the equivalence assumption. The '10' indicates that we are confident, before the study, that the equivalence is at least as high as the equivalence we would get, if out of 22 subjects, 20 are classified the same way with both devices (10 successes on both devices and 10 failures on both devices also). If the results are roughly the same whatever the prior distribution (i.e. prior knowledge on equivalence), then it can be concluded that the data are sufficient to give stable results, whatever the prior. Thus, different experts, expressing different opinions through different prior distributions, would logically agree on the results.

The interim posterior distribution was then used to compute the PredP over the future unobserved data, in the following way (described in Fig. 2 and Appendix 1)[8]:

(i) All possible future data combinations are listed (all pairs of outcomes for each of the future patients). For example, if $k=153$ patients have already been recruited and assessed, then $m=480-k=327$ future patients are contemplated, and there are several million possible outcome tables for these 327 additional patients. Given the data at the first interim analysis (for $k=153$ patients, in Table 2), we would expect the outcome for the next 327 patients to be close to (278, 23, 0, 26) rather than, say, (0, 26, 278, 23) or (100, 85, 85,107), which suggests to take account of the probability of each specific table.

(ii) The probability P_T of each of these new tables, conditional on the prior and the data currently observed, is computed. Although straightforward, the details of the computation are out of scope of this paper. The sum of the probabilities (P_T) of all these new tables is naturally equal to 1.

(iii) For each of the possible new tables, the posterior probability of equivalence P_E based on this possible new table can be calculated, usually by simulation. In our example,

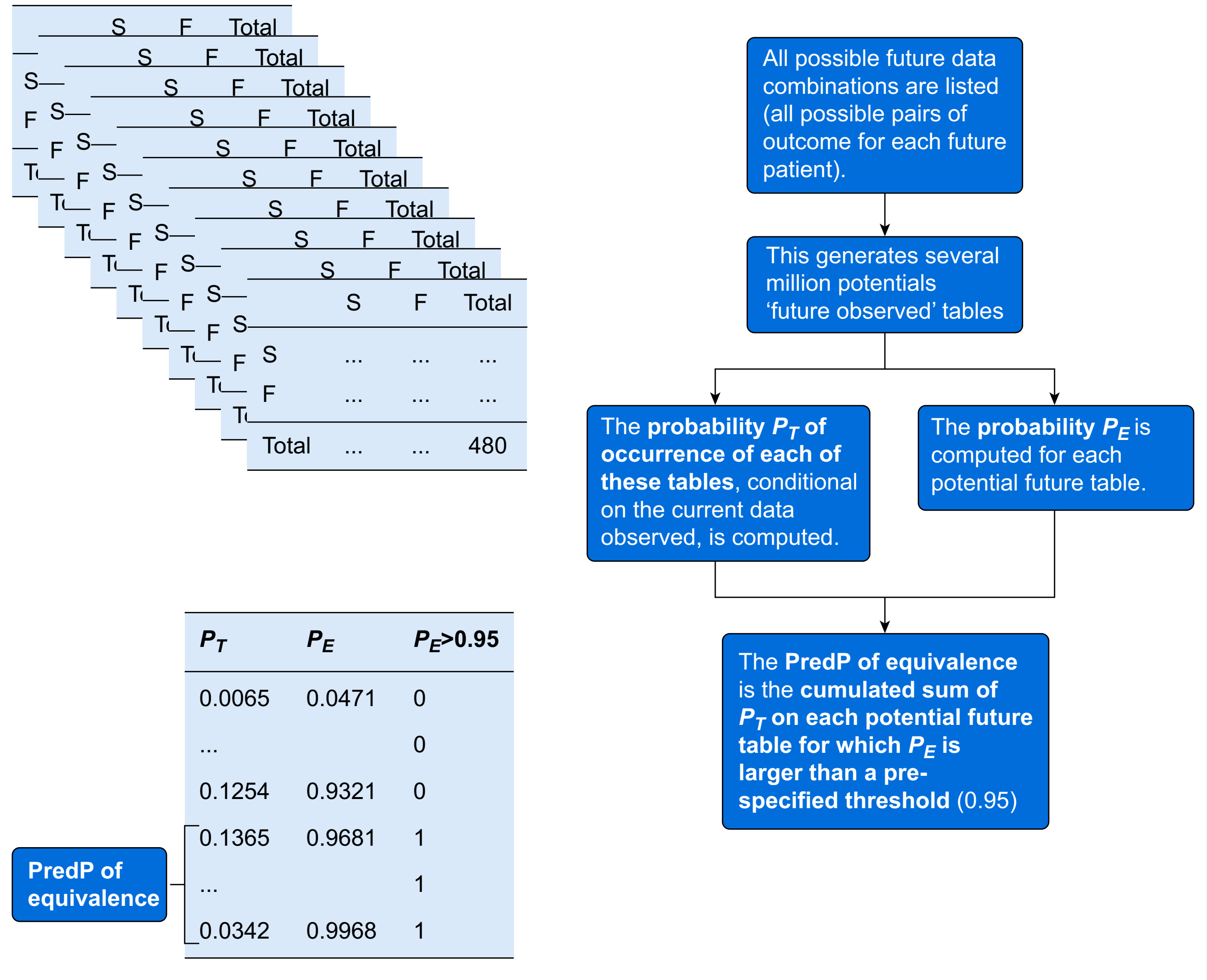
S F Total
S
F
Total
S F Total
S ... ...
F ... ...
Total 480
All possible future data combinations are listed (all possible pairs of outcome for each future patient).
This generates several million potentials 'future observed' tables
The probability P_T of occurrence of each of these tables, conditional on the current data observed, is computed.
The probability P_E is computed for each potential future table.
The PredP of equivalence is the cumulated sum of P_T on each potential future table for which P_E is larger than a pre-specified threshold (0.95)
P_T P_E P_E>0.95
0.0065 0.0471 0
... 0
0.1254 0.9321 0
0.1365 0.9681 1
... 1
0.0342 0.9968 1
PredP of equivalence

Fig 2. Diagram describing the computation of predictive probabilities. In the 'future observed' tables, S is success and F is failure. P_E, posterior probability of equivalence for each table; PredP, predictive probability; P_T, probability of occurrence of each of these tables.

P_E is the posterior probability that the true value of $P_{10}+P_{01}$ is less than 10% for this conjectured new table.

(iv) The decision rule is to claim equivalence if, for any new table, P_E is greater than 0.95. All possible tables with P_E >0.95 are kept and those with P_E <0.95 are discarded. The sum of probabilities P_T of tables for which P_E >0.95 is the predicted probability of concluding equivalence at the end of the study.

The PredP of equivalence is the probability of concluding equivalence at the end of the study, considering all possible future data. It may allow early termination of the trial because of futility or efficacy. A stopping rule is defined before the start of the study. It is predefined by a lower limit (for futility) and an upper limit (for efficacy). If the PredP of equivalence is above or below these thresholds, the study may be stopped. If this probability falls between these two thresholds, the inclusions must continue until the next equivalence PredP calculation is performed with which this interpretation rule will be reused.[13,14] In our study, the lower and upper bounds were set at 0.10 for futility and 0.99 for efficacy. Those thresholds were motivated by the low risk of the study for the patients, a quick patient enrolment, and the need for high level of certainty for this very common medical procedure. PostP and PredP were computed using the 'two patients' Di(0.5, 0.5, 0.5, 0.5) prior and by increasing the number of subjects included (the number of patients was retrospectively and arbitrarily chosen) until PredP exceeded the efficacy or the futility bound.

An interim analysis could have occurred, for instance, after 153, or after 217, or after 352, or after 409 patients had been assessed or any other number of patients, allowing equivalence or futility to be declared, or inclusion to be continued until the next interim analysis or full sample size is reached.

Results

This distribution of success and failure in the R and LB groups (with 153 patients) is presented in Table 2. The predefined threshold of PostP (0.95) was reached during the first interim analysis with 153 patients (a=131, b=11, c=0, and d=11). But, the PredP based on the same data, predicting the results on the

Table 2 Description of result distribution at the time of predictive probability computation (*n*=153). CL, Cormack and Lehane grade; LB, Laryngobloc; R, control.

Laryngoscopy		LB group	
		Success CL 1–2	Failure CL 3–4
R group	Success CL 1–2	131	11
	Failure CL 3–4	0	11

complete sample of 480 subjects, was only 0.675 (Table 3). Computations, with the prior Di(0.5, 0.5, 0.5, 0.5) of PostP and of PredP for the set of arbitrarily chosen numbers of subjects included computed, are described in Table 3. The predefined bound of efficacy of PredP (0.99) was reached with 409 patients.

Considering the sensitivity analysis, whatever selected prior, we could not conclude with a sufficiently high probability that the upper efficacy bound of 0.99 for the PredP would be met on the primary outcome by the end of the study with only 153 or 217 patients. The results of the sensibility analysis done by modifying the prior distribution parameters are described in Table 3. The PredP increased with the rise of the sample size, and it increased more rapidly with the optimistic prior (Di[10, 1, 1, 10]). The study could be stopped before 352 subjects when using this optimistic prior, whilst a larger sample size would be required before stopping accrual if one uses a neutral prior (Di[1, 1, 1, 1]).

Discussion

According to the interim analyses described, we concluded that the study could be stopped somewhere between 352 and 409 patients. Indeed, as the PredP exceeded our efficacy bound in this range of sample size, the PostP predicted to be above the predefined bound of equivalence by the end of the study. It would thus be useless to continue the accrual, and stopping the study should be considered.

If the PostP of equivalence of 0.95 had been considered alone or if a frequentist analysis had been made on the basis of 217 out of 480 subjects, the study would have been prematurely stopped for equivalence. The PredP indicated that this result was nonetheless unstable and that the equivalence was, thus far, not guaranteed on the target sample size. The observed equivalence may be attributable to a random variation that had only a small probability of being confirmed on the final sample.

The sensitivity analysis showed that the PredP increased with increasing sample size, and it increased more rapidly with the optimistic prior (Di[10, 1, 1, 10]) favouring the hypothesis of equivalence. It thus showed that an informative optimistic prior might have further reduced the effective sample size than using the original prior. In case of prior information, the gain in sample size can be substantial.

PredP is effective and a flexible solution for interim analysis of clinical trials. Our example illustrated this use of the PredP for clinical trial monitoring. PredP provides a way to monitor the probability that a trial will be conclusive (or not).

PredP has several advantages for interim analyses of clinical trials. One of the most appealing aspects of PredP is that it allows for early stopping of a trial that shows either a very efficient or a very inefficient device or drug. This does not mean that any trial using PredP will systematically have a lower sample size than the same trial run without PredP, but it may potentially allow for this possibility. In contrast, if PredP suggests that more patients need to be enrolled, then it may be far easier to enrol in a current study than to add enrolment after the maximum enrolment has completed and after the data are analysed. This feature is particularly attractive, considering the economic aspect of a trial management and the growing pressure to finish studies as quickly as possible, in competition with other centres. Finally, the ethical issue may be the most important one in trials involving new drugs or devices in a population of sick patients.

The use of PredP can be summarised as follows: if the PostP of the outcome is high and the PredP is also high, then the data can be considered as ‘stable’, the trial is conclusive and positive, and there is no need to continue the trial. If, in contrast, the PostP of the outcome is high but the PredP is not large enough, this means that the current evidences are poor, the data are ‘unstable’, and the trial must be continued to the next PredP computation. Our work demonstrates the particular interest of the use of the PredP, where the PostP of equivalence between the LB and R groups is rapidly high (0.999), suggesting clearly a trend, with a PredP that reached the efficacy bound in the range of 352 and 409 included patients. To make a long story short, the PredP can be seen as an index of the long-run stability of the PostP.

Conclusions

Using the PredP in the course of the trial may play the role of an internal reproducibility check and can be used as a tool in the current debate on the reproducibility crisis.[15,16] The interim PredP value must be interpreted for what it is: a prediction of the future, and the better the prediction, the better the trial. If the PredP is high and if the final PostP, computed on the complete sample, is high, then the results can be considered, loosely, as reproducible. It is not as strong an argument as an independent confirmative trial, but it is nevertheless a positive argument. PredP can be used with any type of data,

Table 3 Description of interim analyses, which could have occurred if accrual was allowed to continue to the target value. The posterior probability of equivalence (PostP) and the predictive probability (PredP) of declaring equivalence are specified according to various prior. Di, Dirichlet distribution.

Number of subjects included	Patient distribution	PostP from a Di(0.5, 0.5, 0.5, 0.5)	PredP from a Di(0.5, 0.5, 0.5, 0.5)	PredP from a Di(1, 1, 1, 1)	PredP from a Di(10, 1, 1, 10)
153	131, 11, 0, 11	0.999	0.675	0.676	0.830
217	189, 15, 0, 13	0.999	0.776	0.777	0.893
352	311, 23, 1, 17	0.999	0.974	0.974	0.992
409	365, 26, 1, 17	0.999	0.996	0.995	0.999

even though it is easier to apply on qualitative data than on continuous or survival data. Whatever the context (superiority, non-inferiority, or equivalence, for independent or paired data), a Bayesian approach using PredP can be a useful approach to monitoring of a clinical trial.

Authors' contributions

Statistical analysis: DF, NM
Intellectual content: DF, NM
Writing/review of paper: DF, P-OL, PD, EN, KDT, NM
Approval of paper: all authors
All authors agree to be accountable of all aspects of the work.

Acknowledgements

The authors would like to express their thanks to the reviewers for the very pertinent remarks they made, which enabled the authors to improve their draft manuscript substantially.

Declarations of interest

The authors declare that they have no conflicts of interest.

References

1. DeMets DL, Fost N, Powers M. An Institutional Review Board dilemma: responsible for safety monitoring but not in control. *Clin Trial.* 2006; **3**: 142–8
2. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol* 2016; **31**: 337–50
3. Berkson J. Some difficulties of interpretation encountered in the application of the chi-square test. *J Am Stat Assoc* 1938; **33**: 526–36
4. Berger JO. *Statistical decision theory and bayesian analysis.* New York, Springer-Verlag, New York: Springer Science+Business Media; 2013
5. Berry DA. Interim analysis in clinical trials: the role of the likelihood principle. *Am Stat* 1987; **41**: 117–22
6. Ferreira D, Barthoulot M, Pottecher J, Torp KD, Diemunsch P, Meyer N. A consensus checklist to help clinicians interpret clinical trial results analysed by Bayesian methods. *Br J Anaesth* 2020; **125**: 208–15
7. Ferreira D, Barthoulot M, Pottecher J, Torp KD, Diemunsch P, Meyer N. Theory and practical use of Bayesian methods in interpreting clinical trial data: a narrative review. *Br J Anaesth* 2020; **125**: 201–7
8. Lee JJ, Liu DD. A predictive probability design for phase II cancer clinical trials. *Clin Trial.* 2008; **5**: 93–106
9. Campbell G. The experience in the FDA's center for devices and radiological health with bayesian strategies. *Clin Trial.* 2005; **2**: 359–63. discussion 364–78
10. Ferreira D, Vivot A, Diemunsch P, Meyer N. Bayesian analysis from phase III trials was underused and poorly reported: a systematic review. *J Clin Epidemiol* 2020; **123**: 107–13
11. Liu J, Hsueh H, Hsieh E, Chen JJ. Tests for equivalence or non-inferiority for paired binary data. *Stat Med* 2002; **21**: 231–45
12. Inoue LYT, Berry DA, Parmigiani G. Relationship between Bayesian and frequentist sample size determination. *Am Stat* 2005; **59**: 79–87
13. Berry DA. Introduction to Bayesian methods III: use and interpretation of Bayesian tools in design and analysis. *Clin Trial.* 2005; **2**: 295–300. discussion 301–4, 364–78
14. Berry SM, Carlin BP, Lee JJ, Muller P. *Bayesian adaptive methods for clinical trials.* 1st Edn. Boca Raton, FL: CRC Press; 2010
15. Baker M. 1,500 Scientists lift the lid on reproducibility. *Nature* 2016; **533**: 452–4
16. Ioannidis JP. Why most published research findings are false. *PLoS Med* 2005; **2**: e124

Appendix 1. Pseudo code for predictive probability computation

```
list all possible future tables
calculate the probability $P_T$ of each of these tables,
knowing the data already observed and the a priori knowledge of each table;
on each table, calculate $P_E$, the probability of equivalence by the Markov chain Monte Carlo method (i.e. by simulation, the predictive probability is the sum of the probabilities of the tables for which equivalence is concluded: PredP=$P_T$*1, where 1 is the indicator function (1 if $P_E$ >0.95 else equals 0)
    if PredP > efficiency threshold: STOP;
    if PredP < futility threshold: STOP;
else continue inclusions until the next PredP calculation.
```

Handling editor: Paul Myles