HAL open science

Theory and practical use of Bayesian methods in interpreting clinical trial data: a narrative review

David Ferreira, Mael Barthoulot, Julien Pottecher, Klaus D. Torp, Pierre Diemunsch, Nicolas Meyer

To cite this version:

David Ferreira, Mael Barthoulot, Julien Pottecher, Klaus D. Torp, Pierre Diemunsch, et al.. Theory and practical use of Bayesian methods in interpreting clinical trial data: a narrative review. British Journal of Anaesthesia, 2020, 125, pp.201 - 207. 10.1016/j.bja.2020.04.092 . hal-03492185

HAL Id: hal-03492185

https://hal.science/hal-03492185v1

Submitted on 22 Aug 2022

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.

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Version of Record: https://www.sciencedirect.com/science/article/pii/S000709122030369X
Manuscript_2f74703deba2f87b8a0906f413fedf52

Theory and practical use of Bayesian methods in interpreting clinical trial data: a narrative review.

David Ferreira[1,2], Mael Barthoulot[3], Julien Pottecher[4], Klaus D. Torp[5], Pierre Diemunsch[4], Nicolas Meyer[2,6]

[1]Anesthesiology and Intensive Care Department, CHU Besançon, F-25000 Besançon, France. dferreira@chu-besancon.fr.

[2]Université de Strasbourg, iCUBE, UMR7357, F-67412 Illkirch Cedex, France

[3] Institut Pasteur de Lille, Unité d'Epidémiologie et de Santé Publique, Inserm-U1167, Lille.

[4]CHU de Strasbourg, Anesthesiology and Intensive Care Department, IHU-Strasbourg, F-67091 Strasbourg, France

[5]Department of Anesthesiology, Mayo Clinic, Jacksonville, Florida

[6]CHU de Strasbourg, Public Health Department, GMRC, F-67091 Strasbourg, France

Corresponding Author: Dr David Ferreira. Département d'Anesthésie et de Reanimation. CHU Jean Minjoz. 3 Boulevard A. Fleming. 25030 Besançon Cedex. France.

Phone: +33 3.81.66.91.31 fax: +33 3.81.66.90.13

email: dferreira@chu-besancon.fr

© 2020 published by Elsevier. This manuscript is made available under the CC BY NC user license
https://creativecommons.org/licenses/by-nc/4.0/

Running title: Bayesian vs Frequentist methods in trial analysis.

EDITORS KEYPOINTS

- Bayes' theorem was devised by the 18th-century English theologian and mathematician Thomas Bayes.
- The Bayesian view of probability is related to degree of belief, reflecting the plausibility of an event given incomplete knowledge.
- The Bayesian approach uses probabilities for both hypotheses and data.
- This statistical approach is increasingly used in clinical trials, genetic analyses and for applied machine learning.

Abstract

The critical reading of scientific articles is necessary for the daily practice of evidence-based medicine. Rigorous comprehension of statistical methods is essential, as reflected by the extensive use of statistics in the biomedical literature. In contrast to the customary frequentist approach, which never uses uses or gives the probability of a hypothesis, Bayesian theory uses probabilities for both hypotheses and data. This statistical approach is increasingly used for analyses of clinical trial data and for applied machine learning. The aim of this review is to compare general Bayesian concepts with frequentist methods, in order to facilitate a better understanding of Bayesian theory for readers who are not familiar with this approach. The review is intended to be used in combination with a checklist we have devised for reading reports analysed by Bayesian methods. We compare and contrast the different approaches of Bayesian versus frequentist statistical methods by considering data from a clinical trial that lends itself to this comparative approach.

Keywords: Bayes; frequentist; RCT; clinician; theory; methods

Introduction

In order to keep up to date, physicians frequently consider the results of clinical trials, particularly phase III trials. To appropriately interpret trial results, clinicians must understand the statistical methods reported. Currently, two kinds of methods coexist in inferential statistics: frequentist and Bayesian methods. Frequentist statistics are predominant in the field of biomedical research, based on the Null Hypothesis Test (NHT). The probability of an event is defined as the long-term frequency of occurrence of this event, in a series of repeated trials or in a set of "identically" conducted experiments. The name frequentist statistics is derived from this definition of probability. This probability is empirical and deemed objective because it relies on past observed data only.

However, recent advances in the implementation of older Bayesian statistical approaches [1 2] have led to a renewed interest in Bayesian methods, where probability is a measure of the degree of confidence or knowledge (or belief) in the occurrence of an event. This definition is consistent with the meaning of probability in everyday language. Bayesian probability is subjective and relates to statement on the credibility of an event. In this approach, the parameter of interest, with unknown values, is used in a probability distribution (the set of values that the parameter can take, with their probabilities of occurrence). Bayesian probability distributions express our degree of knowledge of the parameter with *prior* probabilities (knowledge of the parameter before further study), *posterior* probabilities (conditional on study data) and predictive probabilities (relating to data yet to be observed).[3]

Many studies have now demonstrated the feasibility and relevance of these statistical methods in clinical trials [4–8]. Consequently, an increasing number of therapeutic trials with results analysed by Bayesian methods are being published in major journals [9–20]. Recommendations for Bayesian analyses have been developed [21–23], but these were primarily developed for researchers. They are not aimed at readers unfamiliar with Bayesian methods. For this reason, we have also developed a tool destined for practitioner to aid them in the understanding of clinical trials with analyses described by Bayesian terminology.[24] To further facilitate understanding of the Bayesian approach, here we compare Bayesian and frequentist approaches using the IMMERSION clinical trial as an example.[25]

Methods

Clinical trial

We apply the Bayesian approach to a clinical trial following the IMRAD structure [26]. IMMERSION was an open prospective randomised controlled study with parallel groups. Non-pregnant women were assigned to either water immersion (2 hours in a bathtub ((Bath group)) or to bed rest for the same duration at neutral temperature (Bed group). Diuresis (primary endpoint) was assessed by measuring voiding volume. A Bayesian statistical analysis was performed. The mean difference and its 95% credibility interval, posteriors probabilities were computed for the main outcome. The main analytical objective was to determine the difference in diuresis levels, hereafter called θ, between the intervention group (partial immersion: "bath") and control group ("bed rest"). The required sample size was 20 subjects per group for an expected mean diuresis difference of 100 mL, with a standard deviation of 100 mL, a type I error rate of 5%, a type II error rate of 20% in an equitail test [25].

Part 1. Theory and general concepts of frequentist and Bayesian methods

Study objective defined by probability

Considering a toin coss, the frequentist approach dictates that if one tosses a coin 10 times and for example, gets 6 tails, then the probability of tails is, according to this experiment, 60%. By contrast, in Bayesian theory, for a coin assumed to be fair, the prior probability of heads or tails is 50%. If, for example, one gets tails on 6 of 10 tosses, then the posterior probability of tails, according to the results of this experiment, will be a combination of the prior probability of 50% and of the observed 6/10 tail tosses. Therefore, this view distinguishes the notion of frequency observed during the experiment (6/10) from that of the estimated value of the unknown probability of getting heads, in a population of total tosses. The final posterior estimated value of the unknown probability of heads can thus be close to the prior 50% probability hypothesis or close to the observed 60% probability depending on the relative weigth of the prior and of the data.

Hypothesis.

To compare two means, the frequentist methodology asks the clinician to *define two hypotheses*:
H0: there is no difference between the two groups.

H1: there is a difference between the two groups.
The magnitude of the difference is only specified for calculating the number of required subjects but is not the formal subject of the test.

In the IMMERSION trial, the null hypothesis is H0 i.e. there is no difference between the mean diuresis of the IMMERSION group and the mean diuresis of the control group. Alternative hypothesis is H1 i.e. there is a difference between the two groups in terms of diuresis.

The Bayesian methodology asks the clinician to *define the prior knowledge of plausible means of each group*, which the statistician translates into a probability distribution (corresponding to a description of all possible values of the estimated mean and their probabilities of occurrence), called a *prior* distribution to summarise this knowledge.
In most cases, partial knowledge of the difference of the means (θ) is available before the data are collected. This prior information may come from previous similar experiments, expert opinions of the phenomenon or basic physiological knowledge.

In the IMMERSION trial, authors assume that the hourly diuresis for a healthy woman is in most cases (i.e., with a very high probability) between 0 and 1 L. It is never negative or > 4 L. The starting hypothesis is that the diuresis of the experimental group is 50 ml greater than that of the control group. Therefore, we can start from a hypothesis of a mean diuresis level of 100 $ml.h^{-1}$ in the control group and 150 $ml.h^{-1}$ in the experimental group with identical standard deviations of approximately 40 ml.

Theoretical framework (table 1).

In the NHT approach, only H0 is formally tested, and the whole process of the test is carried out under the consideration that this hypothesis is true. However, the process attempts to show that it is false. This hypothesis test establishes a decision rule that will lead us to consider non-rejection or rejection of the null hypothesis H0. Since this decision is based on the results of a sample that is only a part of the population, we cannot make a decision with certainty, and therefore a risk of error must be considered. The decision to accept or reject H0 is made so as to minimise the risk of a wrong decision in a hypothetical series of test repetitions.

By contrast, Bayesian theory does not use test in this way. The Bayes theorem allows for determining probabilities, as probability of occurrence of a parameter of interest θ (in IMMERSION, probability that the difference in mean diuresis of the immersion group is

greater than that of the control group) knowing the data observed in the experiment (termed y) (*i.e.,* here measurements of diuresis values in each participant in both groups). If we call θ the difference in mean diuresis, then according to the Bayes theorem:

$$\Pr(\theta|y) = \Pr(y|\theta)\ \Pr(\theta)\ /\ \Pr(y) \qquad (1)$$

Each term of the theorem has a usual denomination. The explanation of these terms will be presented in table 1. The term Pr (θ) is the *prior* probability of θ. The term Pr (y | θ) is the likelihood function of θ. The term Pr(θ|y) is the *posterior* probability.

Concept of estimation.

In the frequentist approach, the parameter of interest θ is unknown but is considered constant. θ is estimated considering that everything we know about θ comes from the data. Therefore, this estimation relies solely on the likelihood and probability of the data under H0. H0 has no probability in itself.

In the IMMERSION study, the Bayesian approach reports a primary outcome measure comparing the mean diuresis between the two study groups. Previous knowledge of an expected mean difference in diuresis of 100 mL between the groups, under the assumption of a standard deviation of 100 mL diuresis in each group, was only used in the sample size determination.

The principle of Bayesian estimation is to consider the parameter θ as unknown. What is unknown is uncertain and is thus given a probability distribution (see Bayesian definition of probability). We then estimate the probability that θ is within a certain range of values. To estimate θ consists in adjusting the *prior* knowledge on θ using the information provided by the data for the experiment, through the likelihood. We then examine the conditional distribution of θ knowing the data y, *i.e.* the *posterior* distribution.

In the IMMERSION study, the chosen *prior* distribution for mean diuresis levels (a normally-distributed outcome) in each group was a normal, or Gaussian, distribution, with two parameters: the mean and the standard deviation. The prior was N(m = 2.68, sd = 1) in the Bath group and N(m = 1.75, sd = 1.56) in the bed group, based on the results of Katz [36], expressed in mL standardised on the mean study participant weight (59.9 kg).

Bayesian and frequentist estimators can be numerically very close, especially when there are many data points. When there are few data points, the difference may be great and depends on

the choice of the *prior* distribution and we could compensate for this lack of information with a *prior* knowledge. The more observations we have, the more the relative importance of the *prior* information decreases.

Use of data.

For frequentist, in analysing a clinical trial, the previous data are not used explicitly. They are used for computing the sample size, but they do not appear in the final analysis of the data. The clinician must define an expected mean difference and standard deviation to calculate the sample size of the study. In the context of a meta-analysis, previous data are formally included in the computation, but only as "pure data" and not as an accumulating knowledge.

For Bayesians, in analysing a clinical trial, the previous data can be used explicitely, by introducing them in the prior distribution. Considered as a prior knowledge, they are thus formally included in the estimation process and mixed with the observed data. In the context of a meta-analysis, previous data are included in the estimation process through the prior and the Bayesian meta-analysis can reach conclusion much sooner than its classical counterpart that consider the study as independent [28].

Results: Practical application of statistical analyses

Study objective defined by probability.

Frequentist: a single question using the null hypothesis test is addressed, namely that there is no difference in diuresis reduction after partial immersion (bath) versus after bed rest.

Bayesian: the main objective is determined by asking a single question in a probabilistic form (i.e., as a degree of belief). In IMMERSION, the main objective would be to determine what is the probability that diuresis is reduced by at least x mL after partial immersion (bath) versus after bed rest.

Process of data analyses

The classical test is performed as follows:

1) The null and alternative hypotheses are expressed, in particular by specifying the targeted effect size, based on previous parameter estimations, considered as known, or as hypothetical values of clinical interest. The hypotheses of the IMMERSION study are defined in hypothesis section.

2) Determination of the specific test or model to use in accordance with the nature of the variable (t-test for a quantitative outcome, etc.) and the conditions of application of the test. The main objective of the IMMERSION study is to compare a quantitative variable (diuresis) between the interventional group (bath) and the control group (bed rest) in a unilateral situation (higher diuresis level in the bath group). If the test assumptions are met (normality of distributions and homogeneity of variances), the test to be used is the Student's *t* test. If the conditions are not met, a non-parametric Mann and Whitney test will be used instead.

3) Determination of the test values, after defining the acceptance and rejection zones based on the type I rate and the power of the test. In our example, we use the *t* test. The *t* test threshold value delineating the acceptance and rejection zones was defined with an α risk of 5% for n=40.

4) Calculation of the test value from the sample data and conclusion of the test. The *t* test value is calculated and compared to the t-test threshold value at a, say, 5% level. Similarly, a p-value can be computed and compared to the alpha level.

According to Gelman [29], the process of Bayesian data analysis can be described by dividing it into three steps:

1) A probability distribution is determined for all variables of the studied problem. The sources used to construct the *prior* distribution can be derived from a meta-analysis, previous studies [5 10], expert opinion [5 30] or a biophysical theory. However, the use of expert opinion is debated because this introduces a more subjective nature into the study outcome analysis.

The process of the expert expressing knowledge and formulating it mathematically in the *prior* distribution is called *elicitation*.

The model should be consistent with what is known about the scientific problem involved and the nature of the parameters involved in the analysis (mean, proportions, etc). The prior distribution specifies what is known about the difference of interest. Clinically relevant differences to be tested are specified before the study.

2. From the observed data and the *prior* distribution, the *posterior* distribution is calculated by simulation (using software such as BUGS, JAGS...). The main objective of the IMMERSION study is the same as in a classical framework.

3. Evaluation of the goodness of fit of the model and the implications on the resulting posterior distribution. Using the suitable model, software and technic, for the main outcome, the mean difference (in favor of the bath group) and its 95% credible interval (CrI) are computed thanks to the posterior distribution, as well as the probability that the difference is positive and that the difference is > 0.835 $mL.kg^{-1}.h^{-1}$, which corresponds to a diuresis difference of 50 $mL.h^{-1}$, standardised on the mean study participant weight (59.9 kg).

Presentation of the results

A frequentist result usually provides the estimated parameter of interest with its 95% confidence interval (CI). When using a statistical test in addition to the test results and its 95% CI, the p value is computed.

In the IMMERSION study, the difference between the bath and bed rest group hourly diuresis was 1.23 $mL.kg^{-1}.h^{-1}$ (95% CI 0.42; 2.05) p=0.0039. The p value means that under the null hypothesis (typically that the intervention has no effect), there is a 0.39% chance of observing a mean diuresis gap as large as that observed in the study.

The *posterior* distribution is the main result of a Bayesian analysis and it encompass all the values that the parameter of interest can take *a posteriori*. From this distribution, one can deduce a mean or median with a "range" for the parameter of interest, called a Credible Interval (CrI). A Bayesian result is thus obtained by **estimating the posterior value of the parameter of interest with its 95% CrI**.

In the IMMERSION study, the *posterior* probability that the diuresis difference is at least 0.835 $mL.kg^{-1}.h^{-1}$ in bath *vs*. bed rest (considering an informative prior from Katz et al N(2.68,1))[36] was estimated to be 0.782, with a mean difference of 1.26 $mL.kg^{-1}.h^{-1}$ (95% CrI 0.20; 2.32). It was positive, so hourly diuresis was larger for the bath than bed rest group. The probability of this difference being positive was 0.99.

Interpretation of the results

The frequentist situation corresponds to a deductive inference, which begins with a hypothesis about the world and tests whether the observations are consistent with this hypothesis [31].

We see that the p value is not the probability that H0 is right or false. It does not give any indication of the magnitude of the difference in the population, which is the criterion of clinical interest. It is in fact confounded with the difference magnitude and the sample size.

Moreover, the fact that the threshold of significance is set at 0.05 is totally arbitrary [13 14] and completely ignores the clinical-biological likelihood and previous knowledge [15].

For Bayesians, different thresholds can be tested on the same *posterior* distribution without having to correct the tests for multiple comparisons. For example, the probabilities that the diuresis difference is greater than, say, 0.5, 0.7 or 1 $mL.kg^{-1}.h^{-1}$ can be computed and compared without considering the number of comparisons made because they all derive from the same distribution.

The estimation of the value of the unknown parameter is based on a random sample. This estimate will result from the combination of the known information on the parameter before the experiment and the data resulting from the experiment. The goal is not to estimate the mean difference in a given sample (this is of no interest) but to make a general estimate of this difference for the population of interest, based on the observation provided by a single sample. The use of the concept of subjective probability makes it possible to really calculate the probability, in the population, that the parameter of interest is within a given range, based on a single sample. The effect magnitude can thus be "isolated" from the sample size effect.

Interpretation of intervals

Common statististical analyses rely on both descriptive and inferential analyses. The descriptive analysis is done giving the point estimation (mean, proportion, quartiles, etc) while the inferential analysis relies on formal testing of the parameter. These two approaches are separated in the classical statistical context while in Bayesian statistics they are both computed on a unique object, i.e. the posterior distribution. In both paradigms, the point estimate is given with an interval: confidence interval in the frequentist method and credibility interval in the Bayesian world.

The 95% CI means that if an experiment was repeated an infinite number of times under the same conditions, 95% of the estimated intervals would contain the true value of the parameter, whose value remains unknown [32 33]. It is a description of what value a parameter can take under repeated sampling. Contrary to what intuition may make one believe, it does not provide the probability that the value is within a given interval.

The 95% CrI indicates that there is a probability of 0.95 ("95% chance") that the true value of the parameter is within the interval. The 95% *posterior* CrI depends in part on the *prior* distribution (credible values taken on diuresis levels before the study are performed) and in

part on the observed data (diuresis values during the study). When the initial information is vague (lowly or uninformative *prior*), the 95% CrI, in its usual version, will have approximately the same bounds as the classical 95% CI.

Initial *prior* information that is relatively precise will reduce the dispersion of the *posterior* distribution and thus reduce the 95% CrI, making it more precise. Moreover, one can calculate the probability that the parameter of interest value is greater or less than a threshold (set before the study) or whether it is in a predefined given interval.

Sensitivity analyses

Sensitivity analysis is defined as "a method to determine the robustness of an assessment by examining the extent to which results are affected by changes in methods, models, values of unmeasured variables, or assumptions" with the aim of identifying "results that are most dependent on questionable or unsupported assumptions" [34 35].

There are different kind of scenarios [35]:

- Modification of cut-offs or definition of outcomes
- Methods of inclusion of outliers
- Use of missing data or not
- Intention to treat or per protocol analysis etc.

Statistical methods can be modified: for example,

- Parametric and non-parametric methods
- Use of different methods of adjustment (baseline characteristics, the kind of method of adjusting)

All the elements listed above concern both approaches, but the Bayesian approach also requires that assumptions about prior distributions be included in the sensitivity analysis. Indeed, sensitivity analyses are frequently used to test the impact of several *prior* distributions on the estimation of the parameter of interest, and on its *posterior* distribution. They are essential because they allow for checking the result stability under varying initial assumptions.

If the results are essentially identical when different *prior* distributions are used, then we can consider the data to be of sufficient weight and that the conclusion has been reached.

Otherwise, ideally, the number of individuals included in the analysis should be increased until a stable conclusion is reached. However, this is not possible, most of the time, in a fixed design. This calls for tempering the results on the one hand and planning a new study to increase the amount of data available on the other.

Thus, in addition to the "principal" *prior* distribution, complementary analyses using an uninformative *prior* distribution, an "enthusiastic" *prior* distribution (in favor of the hypothesis tested), and a "pessimistic" *prior* distribution (not in favor of the hypothesis tested) can be performed. It is also possible to perform a sensitivity analysis on the model, or the estimation methods used. If different priors yield different results and conclusion, the authors should acknowledge this.

Discussion

This comparison shows that, beyond the differences in their use, the difference in the interpretation of these two methods is also very important. All clinicians naturally have a moderate sense of Bayesian reasoning, often without knowing it. In the search for a diagnosis, the question is "What is the probability that my patient has this or a different pathology?". The clinician intuitively uses a pretest probability for each disease (in the form of a list of diagnoses consistent with early observations, each diagnosis being more or less probable, depending on the frequency of the disease in the general population and the frequency of each symptom for each disease). The pretest probability of each disease under consideration is increased or decreased according to the results of the diagnostic test (resulting in a post-test probability). The same reasoning is used when making decisions about treatments. The initial confidence about whether or not a treatment will work is influenced by the clinician's knowledge of pathophysiology and pharmacology, by reading reports of high-quality RCTs [37], and should be confirmed by practice [38].

The *prior* distribution is called "**informative**" if it conveys a lot of information on the parameter (*i.e.,* is precise on its prior values). "**Non-informative**" or better "**lowly informative"** distribution can also be used: all possible values of the parameter of interest θ are, in the eyes of the expert, equally (or almost equally) likely; that is, he (she) will not bet more on one value than on another. Using this kind of *prior*, it is the data (the likelihood) that will have the greatest impact on the *posterior* distribution [5 39 40]. A "typical" non-informative *prior* distribution is a normal distribution with a zero mean and very large variance [39]. In

phase III trials, non-informative (lowly) informative priors are generally preferred by regulators.

To follow a hypothetical deductive scientific approach, **the *prior* distribution must be defined before the data collection**. The *prior* distribution is susceptible to biases if it is constructed after data collection, especially if it is based on expert opinion. However, the *prior* could be revised during the study if other information becomes available, provided that the revisions occur independently of the results of the study.

In Bayesian methodology, the *prior* knowledge is explicitly formalised, enabling authors to use informative knowledge as soon as possible under the control of the reader's common sense. In the IMMERSION study, assuming that a mean diuresis level is between -10^6 $L.h^{-1}$ and $+10^6$ $L.h^{-1}$ would be difficult. However, these are values considered possible by the alternative hypothesis in the frequentist null hypothesis test. With Bayesian methods, the combination of the plausible values for the parameter of interest before the experiment (*prior)* and the information from the experiment allow for obtaining our result: the *posterior* distribution (all the values that *posterior* parameters can take and their probabilities). When the sample size is large ($n \rightarrow \infty$), the influence of the *prior* distribution fades and the likelihood of the data makes up the bulk of the *posterior* distribution. When the state of knowledge about the problem in question permits only a very vague *prior*, the likelihood of the data will also have a predominant impact on the *posterior* distribution. Therefore, there will be disagreement only when the data are insufficient to yield the same *posterior* distribution for different prior assumptions. If the data are of sufficient quantity, all the different *prior* values will approximately approach the same *posterior* distribution and the debate will be settled. Furthermore, the posterior distribution is often difficult to obtain analytically which requires simulation through specific statistical software. This is one of the reasons why Bayesian methods have sometimes been deemed to be difficult to use in the past.

In frequentist methodology, this prior knowledge is implicitly used for the parameter of interest when calculating the sample size. This calculation can never be reassessed without prior agreement in the protocol. In practice, it undergoes the fluctuations inherent in the daily constraints of research (patient lost to follow-up, protocol violations, etc.), therefore making the conclusions of the study obsolete [41]. The final usable result of the statistical tests is the p value, which determines a threshold of arbitrary significance to reject the null hypothesis. The

p value does not measure the probability that the hypothesis studied is true or that the data are produced by chance alone.

Many clinicians instinctively interpret the p value by using an inductive reasoning that is characteristic of the Bayesian method [31]. In a clinical case, the frequentist statistician asks "What is the probability of having a temperature > 39.5º with a diagnosis of influenza", whereas the Bayesian statistician asks "What is the probability of having the flu, knowing that the temperature is > 39.5º", the typical reasoning used in a diagnostic procedure.

In summary, we have compared Bayesian methods with frequentist methods using the IMRAD structure to provide a better understanding of the general theories and concepts of these two methods. This document is intended to serve as an adjunct to our reading grid [24] but can also be used individually to understand Bayesian methodology.

Acknowledgments

Assistance with the article. We thank Laura Smales (BioMedEditing, Toronto, Canada) for English editing.

Authors' contributions :

The authors (D.F., M.B., J.P., K.T., P.D. and N.M.) contributed equally to the development, drafting and final review of the manuscript.

Financial support and sponsorship: This work was supported by the Centre Hospitalier Universitaire de Strasbourg, France.

Conflicts of interest: none declared.

References

1. Gilks WR TA. A Language and Program for Complex Bayesian Modelling. *J R Stat Soc Ser Stat* 1994; **43**: 169:177

2. MRC Biostatistics Unit. The BUGS Project: Bayesian inference using Gibbs sampling.

3. Ferreira D., Ludes P.O., Diemunsch P, Noll E,Torp KD, Meyer N. Bayesian Predictive Probabilities Monitoring for Clinical Trials: a propos of a Comparison of Two Laryngoscopes. *BJA* Manuscript submitted for publication.

4. Berry. Statistics: A Bayesian Perspective. 1996.

5. Spiegelhalter D. J. AKR. Bayesian Approaches to Clinical Trials and Health-Care Evaluation. John Wiley & Sons, editor. 2004.

6. Goodman SN. Introduction to Bayesian methods I: measuring the strength of evidence. *Clin Trials* 2005; **2**: 282–90; discussion 301-4, 364–78

7. Louis TA. Introduction to Bayesian methods II: fundamental concepts. *Clin Trials* 2005; **2**: 291–4; discussion 301-4, 364–78

8. Berry DA. Introduction to Bayesian methods III: use and interpretation of Bayesian tools in design and analysis. *Clin Trials Lond Engl* 2005; **2**: 295–300; discussion 301-304, 364–78

9. Roberts KA, Dixon-Woods M, Fitzpatrick R, Abrams KR, Jones DR. Factors affecting uptake of childhood immunisation: a Bayesian synthesis of qualitative and quantitative evidence. *Lancet* 2002; **360**: 1596–9

10. Brophy JM, Joseph L. Placing trials in context using Bayesian analysis. GUSTO revisited by Reverend Bayes. *JAMA* 1995; **273**: 871–5

11. Muss HB, Berry DA, Cirrincione CT, et al. Adjuvant chemotherapy in older women with early-stage breast cancer. *N Engl J Med* 2009; **360**: 2055–65

12. Baeten D, Baraliakos X, Braun J, et al. Anti-interleukin-17A monoclonal antibody secukinumab in treatment of ankylosing spondylitis: a randomised, double-blind, placebo-controlled trial. *Lancet* 2013; **382**: 1705–13

13. Shah PL, Slebos DJ, Cardoso PFG, et al. Bronchoscopic lung-volume reduction with Exhale airway stents for emphysema (EASE trial): randomised, sham-controlled, multicentre trial. *The Lancet* 2011; **378**: 997–1005

14. Wilber DJ, Pappone C, Neuzil P, et al. Comparison of antiarrhythmic drug therapy and radiofrequency catheter ablation in patients with paroxysmal atrial fibrillation: a randomized controlled trial. *Jama* 2010; **303**: 333–340

15. Singh SM, Austin PC, Chong A, Alter DA. Coronary angiography following acute myocardial infarction in Ontario, Canada. *Arch Intern Med* 2007; **167**: 808–13

16. Tinetti ME, Baker DI, King M, et al. Effect of dissemination of evidence in reducing injuries from falls. *N Engl J Med* 2008; **359**: 252–61

17. Gausche M, Lewis RJ, Stratton SJ, et al. Effect of out-of-hospital pediatric endotracheal intubation on survival and neurological outcome: a controlled clinical trial. *JAMA* 2000; **283**: 783–90

18. Morris RK, Malin GL, Quinlan-Jones E, et al. Percutaneous vesicoamniotic shunting versus conservative management for fetal lower urinary tract obstruction (PLUTO): a randomised trial. *The Lancet* 2013; **382**: 1496–1506

19. Cannon CP, Shah S, Dansky HM, et al. Safety of anacetrapib in patients with or at high risk for coronary heart disease. *N Engl J Med* 2010; **363**: 2406–15

20. Holmes DR, Teirstein P, Satler L, et al. Sirolimus-eluting stents vs vascular brachytherapy for in-stent restenosis within bare-metal stents: the SISR randomized trial. *Jama* 2006; **295**: 1264–1273

21. Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. Bayesian methods in health technology assessment: a review. *Health Technol Assess* 2000; **4**: 1–130

22. Group BaSiS. Bayesian standards in science (BaSiS).

23. Sung L, Hayden J, Greenberg ML, Koren G, Feldman BM, Tomlinson GA. Seven items were identified for inclusion when reporting a Bayesian analysis of a clinical study. *J Clin Epidemiol* 2005; **58**: 261–8

24. Ferreira D., Barthoulot M., Pottecher J., Torp KD, Diemunsch P., Meyer N. A check-list for Bayesian clinical trials in anesthesiology. *BJA* Manuscript submitted for publication.

25. Noll E., Toramania A., Chamaraux-Tran TN., Charton A., Ferreira D., Meyer N., Pottecher J., Diemunsch P. Prospective,randomised,controlled trial assessing the effects of water immersion, conducted according to parturition hydrotherapy on diuresis stroke volume and preload-responsiveness in healthy volunteers. *EJA* Manuscript submitted for publication.;

26. Uniform Requirements for Manuscripts Submitted to Biomedical Journals. *N Engl J Med* Massachusetts Medical Society; 1997; **336**: 309–16

27. Égré P. Isabelle Drouet (ed.), Le bayésianisme aujourd'hui. Fondements et pratiques. *Œcon Hist Methodol Philos* 2017; 453–8

28. Halsey LG. The reign of the p-value is over: what alternative analyses could we employ to fill the power vacuum? *Biol Lett* 2019; **15**: 20190174

29. Gelman A. Bayesian data analysis. Third edition. Boca Raton: CRC Press; 2014.

30. Parmar MK, Griffiths GO, Spiegelhalter DJ, Souhami RL, Altman DG, van der Scheuren E. Monitoring of large randomised clinical trials: a new approach with Bayesian methods. *Lancet* 2001; **358**: 375–81

31. Wijeysundera DN, Austin PC, Hux JE, Beattie WS, Laupacis A. Bayesian statistical inference enhances the interpretation of contemporary randomized controlled trials. *J Clin Epidemiol* 2009; **62**: 13-21 e5

32. 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

33. Neyman. Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. *Philos Trans R Soc Lond Ser Math Phys Sci* 1937; **236**: 333–80

34. Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiol Drug Saf* 2006; **15**: 291–303

35. Thabane L, Mbuagbaw L, Zhang S, et al. A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. *BMC Med Res Methodol* 2013; **13**: 92

36. Katz VL, Ryder RM, Cefalo RC, Carmichael SC, Goolsby R. A comparison of bed rest and immersion for treating the edema of pregnancy. *Obstet Gynecol* 1990; **75**: 147–51

37. Patel A. Carotid artery stenosis and randomized controlled trials-should we abandon the gold standard. *Neurology* 2014; **10 suppl.1**

38. Wheatley K, Clayton D. Be skeptical about unexpected large apparent treatment effects: the case of an MRC AML12 randomization. *Control Clin Trials* 2003; **24**: 66–70

39. Diamond GA, Forrester JS. Clinical trials and statistical verdicts: probable grounds for appeal. *Ann Intern Med* 1983; **98**: 385–94

40. Berry DA. Bayesian clinical trials. *Nat Rev Drug Discov* 2006; **5**: 27–36

41. Baker M. 1,500 scientists lift the lid on reproducibility. *Nature* 2016; **533**: 452–4

Table 1. Interpretation of theory, according to frequentist and Bayesian methods

Null Hypothesis test (NHT) approach	Bayesian approach
The probabilities of error are known as *alpha* (α) *and beta* (β). The risk α, or type I error rate, is the probability of rejecting the null hypothesis (H0) as false when it is true. The risk β, or type II error rate, is the probability of not rejecting the null hypothesis (H0) when the alternate hypothesis (H1) is true. The complementary probability of the type II error ($1-\beta$) defines the power of the test and represents the probability of rejecting the null hypothesis (H0) when the hypothesis (H1) is true. In other words, the conclusion provided by the NHT does not measure the probability that H0 is true or false, but the probability of a given result in a repetitive process [26]. The p value is the probability, under H0, of obtaining a statistic as extreme as the value observed in the sample. Given a threshold of significance, we compare p and α to decide whether to reject or not reject H0.	The expression (1) can be simplified: $$\textit{Posterior} \propto \textit{Prior} \times \textbf{Likelihood}$$ **("∝" symbol** reads "is proportional to") The *prior distribution* summarises the information available on the parameter of interest before the collection of our data. It corresponds to all possible values that the diuresis in the two groups can plausibly have before the study is carried out (see example in hypothesis section). It may be based on data from previous trials (other cases are described in this paper). It is obtained by the combination of the *prior* distribution (what we know about the parameter before the experiment) and the *likelihood* (what the data tell us about the parameter according to its prior probability). The data are formally turned into accumulating statistical knowledge through the use of Bayesian theorem into the posterior distribution.

- if $p < \alpha$, we reject the null hypothesis H0 (in favor of H1)
- If $p > \alpha$, we do not reject H0 (in favor of H0).

We can then interpret the p value as the smallest threshold of significance for which the null hypothesis is accepted.

Two results are possible:

1) H0 is not rejected, and we admit there is no difference between the mean diuresis levels of the two groups.

2) We reject H0 and accept H1, and we admit there is a difference between the mean diuresis levels of the two groups.

The posterior distribution describes all we know about the parameter after the experiment. It thus provides us with the parameter estimate and its credibility interval.