RESEARCH

Bayes factors for superiority, non-inferiority, and equivalence designs

Don van Ravenzwaaij[1*], Rei Monden[1,2], Jorge N Tendeiro[1] and John P A Ioannidis[3]

[*]Correspondence:
don.van.ravenzwaaij@rug.nl
[1]University of Groningen, Department of Psychology, Grote Kruisstraat 2/1, Heymans Building, 9712 TS Groningen, The Netherlands
Full list of author information is available at the end of the article

Abstract

Background In clinical trials, study designs may focus on assessment of superiority, equivalence, or non-inferiority, of a new medicine or treatment as compared to a control. Typically, evidence in each of these paradigms is quantified with a variant of the null hypothesis significance test. A null hypothesis is assumed (null effect, inferior by a specific amount, inferior by a specific amount *and* superior by a specific amount, for superiority, non-inferiority, and equivalence respectively), after which the probabilities of obtaining data more extreme than those observed under these null hypotheses are quantified by p-values. Although ubiquitous in clinical testing, the null hypothesis significance test can lead to a number of difficulties in interpretation of the results of the statistical evidence. Here, we advocate quantifying evidence instead by means of Bayes factors. *Results* We highlight how to calculate Bayes factors for the three types of study designs and illustrate how one might obtain those Bayes factors in practice with reanalyses of existing studies. *Conclusions* Bayes factors for superiority, non-inferiority, and equivalence designs allow for explicit quantification of evidence in favor of the null hypothesis. They also allow for interim testing without the need to employ explicit corrections for multiple testing.

Keywords: Bayes factors; Clinical trials; Statistical inference; Non-inferiority designs

Background

In clinical trials, study designs may focus on assessment of superiority, equivalence, or non-inferiority of a new medicine or other intervention as compared to some control intervention [1, 2]. Typically, evidence in each of these paradigms is quantified with a variant of the null hypothesis significance test (NHST). A null hypothesis is assumed, after which the probability of obtaining data more extreme than those observed under the null hypothesis is quantified by a p-value. The specific null hypothesis that forms the basis of these tests differs depending on the design. A graphical display of each of the three designs is provided in Figure 1.

The first, and by far most common, type of design is the superiority design (see top two rows). In the superiority design, the null hypothesis is that the true population effect size is exactly zero. The test can typically be conceived as being one-tailed, even though in practice superiority analyses often employ a two-tailed test. In other words, the null hypothesis states that a new medicine or other intervention being tested does not work better than an existing placebo or active control. The first row in Figure 1 provides an example of a superiority design in which the null hypothesis was rejected and the second row in Figure 1 provides an example of a superiority design in which the null hypothesis was not rejected.

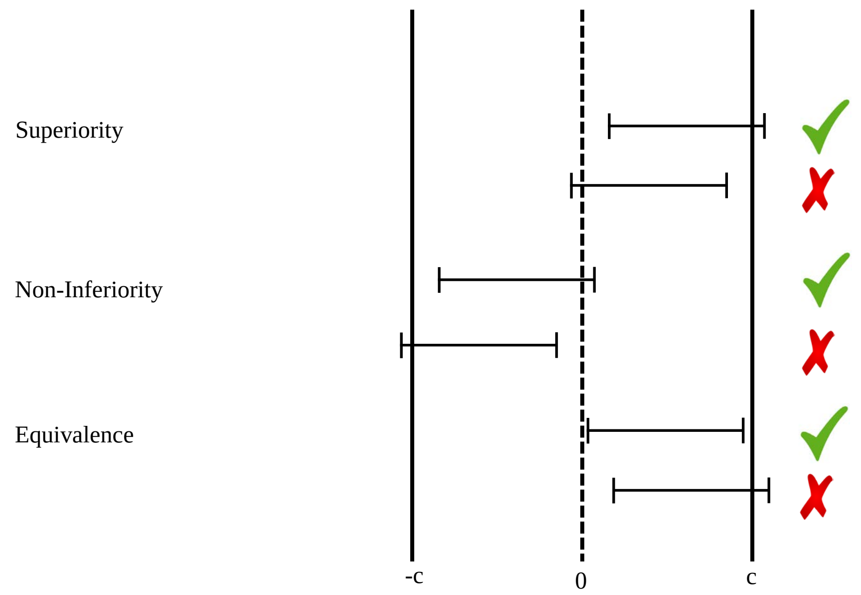

Figure 1 Superiority, non-inferiority, and equivalence designs. Each design type has an example for which the prospective null-hypothesis was rejected (the ones highlighted with check marks) and one for which the null-hypothesis was not rejected (the ones highlighted with fail marks). Error bars indicate 95% confidence intervals. Value $-c$ is the null hypothesis value for non-inferiority testing, values $-c$ and c are the two null hypotheses values for equivalence testing.

The second type of design is the non-inferiority design (see middle two rows). In the non-inferiority design, the null hypothesis is that the true population effect size is lower than $-c$. This amounts to a one-tailed test in which a point-null hypothesis of effect size $= -c$ is compared to an alternative hypothesis of effect size $> -c$. In other words, the relevant test is that a new medicine or other intervention being tested works better than an existing placebo or medication minus an apriori determined amount c. The third row in Figure 1 provides an example of a non-inferiority design in which the null hypothesis was rejected and the fourth row in Figure 1 provides an example of a non-inferiority design in which the null hypothesis was not rejected. Note that it is possible for an intervention to be deemed non-inferior, but simultaneously lower than zero (in this case, the constructed confidence interval would fall between $-c$ and zero in its entirety).

The third type of design is the equivalence design (see bottom two rows). In the equivalence design, one essentially carries out two NHSTs. In this design, the null hypotheses are that the true population effect size is lower than $-c$ *and* higher than c. This amounts to a one-tailed test in which a point-null hypothesis of effect size $= -c$ is compared to an alternative hypothesis of effect size $> -c$ *and* a one-tailed test in which a point-null hypothesis of effect size $= c$ is compared to an alternative hypothesis of effect size $< c$. If both of these null hypotheses are rejected, equivalence is established. Graphically speaking, equivalence is established if the confidence interval falls in its entirety between the borders of $-c$ and c. The fifth row in Figure 1 provides an example of an equivalence design in which both null hypotheses were rejected and the sixth row in Figure 1 provides an example of an equivalence design in which at least one of the two null hypotheses was not rejected. Analogous to non-inferiority designs, it is possible for an intervention to be deemed equivalent, but simultaneously different from zero (in this case, the constructed confidence interval would either fall between $-c$ and zero in its entirety or between zero and c in its entirety). Study results can be interpreted very differently depending on whether the original design was a superiority or an equivalence design [3].

Each of these designs seeks to answer important questions. Unfortunately, the NHSTs employed to carry out statistical inference do not allow researchers to quantify evidence in favor of the null hypothesis. The desire to quantify evidence in favor of the null hypothesis is perhaps most relevant in equivalence designs. We quote Greene, Concato, and Feinstein [4], who say: "…Methodological flaws in a systematic review of 88 studies claiming equivalence, published from 1992 to 1996. Equivalence was inappropriately claimed in 67% of them, on the basis of nonsignificant tests for superiority. Fifty-one percent stated equivalence as an aim, but only 23% were designed with a preset margin of equivalence. Only 22% adopted appropriate practice: a predefined aim of equivalence, a preset Δ, consequent sample size determination, and actually testing equivalence." A non-significant p-value (any $p > .05$) can result from (1) the null hypothesis being true or (2) the null hypothesis being false combined with an underpowered trial citeWitteZenker2017 (that is, if we would have collected more data, the results of our inference would have been statistically significant). In medical research, it is important to distinguish between these two scenarios. Quantifying evidence in favor of the null hypothesis potentially leads to a reduction in the waste of scarce research resources, as research into ineffectual interventions can be discontinued [5].

Another problem with NHST emerges when there is multiple testing in interim analyses. In biomedicine, a range of methods exists that are employed to account for sequential testing and interim analyses, and they all basically change the level of statistical significance by asking for more stringent statistical thresholds to reject the null hypotheses when multiple analyses due to sequential testing or interim re-assessments are performed. However, these correction methods are not always applied. Furthermore, the number of participants tested in clinical trials often changes relative to the number decided upon a-priori based on interim analysis results [6]. Both of these practices lead to an overestimation of the evidence in favor of an effect.

Bayesian methods are an alternative to NHST that allow quantification of evidence in favor of the null hypothesis, sequential testing, and comparison of strength of evidence across different studies [7, 8]. Bayesian methods are increasingly considered for more widespread use in clinical trials (see e.g., [9]; for an overview of different fields, see [10]) and their advantages have been argued many times (e.g., [11, 12, 13, 14, 15, 16], but see [17]). Several approaches to carrying out Bayesian inference exist, but for the remainder of this manuscript we will focus on the *Bayes factor* [18, 19]. The Bayes factor allows for explicit quantification of evidence in favor of the null hypothesis, which means that the interpretational pitfalls associated with non-inferiority and equivalence designs naturally disappear.

In the case of equivalance designs, traditional methods require specification of a potentially arbitrary band around zero, even if clear theoretical grounds for the width of this band are lacking. Bayes factors can quantify evidence in favor of a point null hypothesis or in favor of an interval null hypothesis, depending on which one is theoretically appropriate.

Bayes factors also allow for sequential testing without having to correct for multiple testing (see e.g. the simulation results reported in [20]). It is "...entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience" [21]. To put this quote into perspective, NHST has essentially one decision criterion (i.e., $p < \alpha$). As such, if one employs sequential testing, every additional test increases the chance that this criterion is reached, even if the null hypothesis is true. Bayesian testing does not require a fixed n in the sampling plan because the decision criterion is symmetrical. If one were to decide, for instance, to test until the relative evidence for one hypothesis over the other is at least ten, one would stop when the evidence provided by the data is ten over one in favor of the alternative hypothesis *or* ten over one in favor of the null hypothesis, and one would be wrong once for every ten times one were correct. The Bayes factor will provide progressively stronger relative support for the hypothesis that is true when data continues to be collected.

In what follows, we will describe how to implement Bayes factors for the three types of study design mentioned above.

The Bayes factor

Bayesian statisticians use probability distributions to quantify uncertainty or degree of belief about statistical propositions [22, 23]. For a given statistical model, say M, the prior distribution or prior $p(\theta|M)$ for a parameter θ is updated after encountering data y to yield a posterior distribution or posterior $p(\theta|y, M)$. Bayesian statistics can be viewed as a method for the rational updating of beliefs about statistical propositions. Specifically, Bayes' rule combines the prior, what we believe to be true before having seen the data, with the likelihood, what the data tell us we should believe about the data, to obtain the posterior, what we believe to be true after having seen the data:

$$p(\theta|y, M) = \frac{p(\theta|M)p(y|\theta, M)}{p(y|M)} = \frac{prior \times likelihood}{marginal\ likelihood} \tag{1}$$

In this equation, $p(y|M)$ is the marginal likelihood of the data, a constant that does not involve θ. The posterior $p(\theta|y, M)$ is a mathematical product of prior knowledge $p(\theta|M)$ and the information coming from the data $p(y|\theta, M)$; hence, the posterior contains all that we know about θ (under model M) after observing the data y.

A similar Bayesian procedure can be used for hypothesis testing. Consider for example the choice between hypotheses H_0 (the null hypothesis) and H_1 (the alternative hypothesis). Bayes' theorem dictates how the prior probability of H_0, $p(H_0)$, is updated through the data to give the posterior probability of H_0:

$$p(H_0|y) = \frac{p(H_0)p(y|H_0)}{p(H_0)p(y|H_0) + p(H_1)p(y|H_1)} \tag{2}$$

In the same way, one can calculate the posterior probability of H_1, $p(H_1|y)$. These quantities require specification of the null hypothesis H_0 and the alternative hypothesis H_1. A common choice is to specify the hypotheses in terms of effect size [24]. The null hypothesis then becomes $H_0 : \delta = 0$ and the alternative hypothesis becomes $H_1 : \delta \neq 0$ (or, alternatively, $H_1 : \delta < 0$ or $H_1 : \delta > 0$).

The ratio of the posterior probabilities is given by

$$\frac{p(H_1|y)}{p(H_0|y)} = \frac{p(H_1)}{p(H_0)} \times \frac{p(y|H_1)}{p(y|H_0)} \tag{3}$$

which shows that the change from prior odds of the hypotheses $p(H_1)/p(H_0)$ to posterior odds of the hypotheses $p(H_1|y)/p(H_0|y)$ is given by the ratio of marginal likelihoods $p(y|H_1)/p(y|H_0)$, a quantity known as the Bayes factor, or BF [18, 19].

To see how Bayes factors may be obtained for point null hypotheses, it is illustrative to first consider the calculation of a Bayes factor for interval hypotheses. Let H_0 be that the population effect size falls in an interval around zero: $-c < \delta < c$ and let H_1 be that the population effect size does not fall in that interval: $\delta < -c$ or $\delta > c$. We obtain the Bayes factor by calculating $(p(H1|data)/p(H1))/(p(H0)/p(H0|data))$. The smaller one chooses c (and therefore the interval around zero), the more $p(H0|data)/p(H0)$ will dominate in the calculation of the Bayes factor, as $p(H1|data)/p(H1)$ will tend to 1. In the limit of a point null hypothesis, one can get the Bayes factor by calculating $p(H0)/p(H0|data)$, or by evaluating the ratio of the density of the prior and the posterior, evaluated at $\delta = 0$. This way of calculating the Bayes factor for point null hypotheses is known as the Savage-Dickey procedure, see [25] for a mathematical proof.

Bayes factors represent "the primary tool used in Bayesian inference for hypothesis testing and model selection" [26, p. 378]; Bayes factors allow researchers to quantify evidence in favor of the null hypothesis vis à vis the alternative hypothesis. For instance, when a Bayes factor $BF_{10} = 10$, with the subscript meaning the alternative hypothesis over the null hypothesis, the observed data are 10 times more likely to have occurred under the alternative hypothesis than under the null hypothesis. When $BF_{10} = 1/10 = 0.1$, the observed data are ten times more likely to have occurred under the null hypothesis than under the alternative hypothesis. As for interpreting the strength of evidence as quantified by a Bayes factor, an often-used standard is described in [27]. The authors classify a Bayes factor between 1 and 3 (or, conversely, between 1/3 and 1) as 'not worth more than a bare mention', a Bayes factor between 3 and 20 (or, conversely, between 1/20 and 1/3) as 'positive', and a Bayes factor between 20 and 150 (or, conversely, between 1/150 and 1/20) as 'strong'.

Foundational work on choosing appropriate priors for calculating Bayes factors has been done by Jeffreys [18] and the resulting 'default' Bayes factor remains to this day one of the most popular approaches to obtaining Bayes factors. We will briefly describe the default Bayes factor, then discuss more recent extensions to this work [24, 28].

The default Bayes factor and implementations

Jeffreys' [18] work applies to situations where the two hypotheses to be compared break down into a hypothesis that assigns a single value to the parameter of interest and a hypothesis that specifies a range of values to the parameter of interest. In biomedicine, the practical analogue of this is a point null hypothesis that specifies $\delta = 0$, where δ is an effect size parameter, and an alternative hypothesis that may specify $\delta < 0$, $\delta > 0$, or $\delta \neq 0$.

Jeffreys [18] chose a Cauchy prior distribution with location parameter 0 and scale parameter 1 for the effect size δ parameter. This choice was motivated by the fact that it led to a Bayes factor of exactly 1 in case of completely uninformative data, and on the fact that the Bayes factor would tend to infinity or 1/infinity when the data are overwhelmingly informative. Mathematically, this Cauchy prior corresponds to a normal prior with a mean μ_δ of zero and a variance g that itself follows a scaled inverse chi-square distribution with one degree of freedom, in which the variance is integrated out [29, 30]. It is important to note that Jeffreys' choice of prior was largely motivated by practical reasons, he had no philosophical objections to more informed priors.

The impact of Jeffreys' default Bayes factor had been mostly theoretical until quite recently. An online tool was developed to calculate default Bayes factors for diverse t-test designs [24, available at `http://pcl.missouri.edu/bayesfactor`]. This same group also created the BayesFactor package for the statistical freeware program R [31]. An alternative group, focusing more on informative hypothesis testing, developed the Bain package for the statistical freeware program R [32]. Specialized point–and–click computer software was created for the explicit purpose of doing Bayesian analyses [33] which incorporates many features from the BayesFactor and Bain packages.

In recent work, derivations and R code are provided for (among other things) shifting the center of the Cauchy distribution away from zero [28, code may be found at `https://osf.io/bsp6z/`]. The full equation for obtaining the Bayes factor of the alternative hypothesis $\delta \neq 0$ relative to the null hypothesis $\delta = 0$, BF_{10}, modified from Equation 13 by Gronau et al. [28], is given by

$$BF_{10} = \frac{\begin{array}{c} \displaystyle\int_0^\infty (1+ng)^{-\frac{1}{2}} exp\left(-\frac{\mu_\delta^2}{2(\frac{1}{n}+g)}\right)\left(1+\frac{t^2}{(1+ng)(n-1)}\right)^{-\frac{n}{2}} \times \\ \displaystyle\left[\Gamma\left(\frac{n}{2}\right) {}_1F_1\left(\frac{n}{2};\frac{1}{2};\frac{\mu_\delta^2 t^2}{2\left(\frac{1}{n}+g\right)\left[(1+ng)\,(n-1)+t^2\right]}\right)+\right. \\ \displaystyle\left.\frac{\mu_\delta t}{\sqrt{\frac{1}{2}(\frac{1}{n}+g)\left[(1+ng)\,(n-1)+t^2\right]}}\Gamma(\frac{n+1}{2})\right] \times \left[\frac{\frac{r}{\sqrt{2}}}{\Gamma(\frac{1}{2})}g^{-\frac{3}{2}} exp\left(-\frac{r^2}{2g}\right)\right] dg \end{array}}{\Gamma\left(\frac{n}{2}\right)\left(1+\frac{t^2}{n-1}\right)^{-\frac{n}{2}}} \tag{4}$$

where n is the sample size, μ_δ and g are the mean and standard deviation of the original effect size prior distribution, t is the t-test statistic, Γ denotes the Gamma function, ${}_1F_1$ denotes the confluent hypergeometric function, and r denotes the scale parameter of the Cauchy distribution. This expression allows making modifications to the prior distribution, such as increasing (decreasing) the scale parameter r for fields in which high effect sizes are more (less) frequent and shifting the center of the prior distribution away from zero for the implementation of Bayes factors in non-inferiority designs.

Implementation and Results

In the next subsections, we discuss calculating Bayes factors specifically for superiority and equivalence designs (for which the procedure is essentially identical) and non-inferiority designs. We provide worked examples of reanalyses of real data from publications of clinical trials for each of these to highlight the calculation of these Bayes factors, as well as to provide insight into the merits of this approach over more conventional analyses. Annotated code for conducting these reanalyses is available at `https://osf.io/8br5g/`.

Bayes factors for superiority designs

For superiority designs, the null hypothesis is defined as $\delta = 0$. In order to evaluate this null hypothesis, we can use the Cauchy prior distribution for effect size δ, centered on zero. Ample examples of this approach have been reported elsewhere [11, 34]. Here, we will illustrate this approach with a reanalysis of data reported in [35].

Superiority of racemic adrenaline and on-demand inhalation with acute bronchiolitis

In Skjerven et al. [35], the authors examine the comparative efficacy of adrenaline inhalation by means of bronchodilators versus control (saline inhalations). Specifically, they test for superiority of racemic adrenaline over inhaled saline. In a separate hypothesis, the authors examine whether administration on a fixed schedule is superior to administration on demand. In both cases, the primary outcome is the length of stay in the hospital in hours. The authors conclude that "In the treatment of acute bronchiolitis in infants, inhaled racemic adrenaline is not more effective than inhaled saline. However, the strategy of inhalation on demand appears to be superior to that of inhalation on a fixed schedule." The authors support their first conclusion with a p-value of .42 and their second conclusion with a p-value of .01. Note that the p-values reported by the authors suggest the performed tests were two-sided, although the study goals are more consistent with a one-sided test. In what follows, we report both a one- and two-sided reanalysis.

The reanalysis for the superiority test of racemic adrenaline over inhaled saline proceeds as follows:

- Obtain the standard error, SE_{treat}, from the 95% confidence interval reported in Table 2: $SE_{treat} = 11/1.966 \approx 5.6$
- Calculate the t-statistic for the null-hypothesis that the difference in estimated length of stay between patients that inhaled racemic adrenaline and patients that inhaled saline is zero: $t = \frac{63.6-68.1}{5.6} = -0.80$ (which yields a two-sided p-value of .42).
- We use Equation 4 to calculate a one-sided Bayes factor quantifying the relative likelihood of the one-sided alternative of superiority, $d < 0$, versus the null hypothesis of no effect, $d = 0$, given the data (BF_{-0}). This leads to $BF_{-0} = 0.24$ (or $BF_{0-} = 4.23$), indicating that the null-hypothesis is over 4

times more likely than the one-sided alternative, given the data. The corresponding Bayes factor for a two-sided test is $BF_{10} = 0.15$ (or $BF_{0-} = 6.64$), indicating that the null-hypothesis is over 6 times more likely than the two-sided alternative, given the data.

These and other superiority Bayes factors can be obtained by providing values for the confidence interval margin, sample size n, and group means to the script: $CI_{mar} = (15.5 - (-6.5))/2$, $n_1 = 203$, $n_2 = 201$, $M_1 = 63.6$, and $M_2 = 68.1$ (further details can be found in the annotated code). This reanalysis corroborates the finding of the original authors, who found no significant difference between racemic adrenaline and inhaled saline. The Bayes factor indicates that the null hypothesis is a little over four times more likely than the one-sided alternative of superiority, given the data.

A similar reanalysis for the superiority of fixed schedule inhalation over inhalation on demand yields a one-sided Bayes factor $BF_{0-} = 31.48$ indicating that the null hypothesis is over 31 times more likely than superiority of fixed schedule inhalation over inhalation on demand, given the data. In the sample data, the trend is actually in the direction indicating superiority of inhalation on demand over fixed schedule inhalation. Given that the one-sided test compares two inappropriate hypotheses, we consider the results of a two-sided test more appropriate here. The Bayes factor in favor of a two-sided alternative, BF_{10}, equals 2.24 (recall that [27] classify Bayes factors lower than 3 as not worth more than a bare mention). This finding tempers the conclusion of the original authors: although the data is slightly more consistent with the two-sided alternative hypothesis than with the null hypothesis, the Bayes factor suggests that the evidence is ambiguous and that more study is needed.

In sum, we have seen that Bayes factors can augment interpretation of the statistical evidence for superiority designs in important ways: we can quantify the strength of evidence of one hypothesis relative to another one; and we can explicitly quantify evidence in favor of the null hypothesis. The latter is particularly important for the evaluation of equivalence designs, to which we now turn.

Bayes factors for equivalence designs

The objective of equivalence designs is to show that "the new treatment is at least as good as (no worse than) the existing treatment" [1]. Under a classical NHST approach, it is not possible to test for equivalence directly (the null hypothesis cannot be confirmed). As a result, equivalence needs to be tested by proxy by constructing a band around $\delta = 0$ of $2c$ and evaluating two null hypotheses: $\delta = -c$ and $\delta = c$.

From a Bayesian perspective, the procedure is similar to that of the procedure for superiority designs. Instead of examining the Bayes factor's strength of evidence in favor of H_1, we now examine the strength of evidence in favor of H_0. This removes the ambiguity associated with the traditional approach to equivalence testing. Examine for instance the example where equivalence was demonstrated in Figure 1 (the fifth row). Equivalence was established, because both $\delta = -c$ and $\delta = c$ are rejected (i.e., the confidence interval lies fully between these two boundaries). However, the confidence interval does not overlap with $\delta = 0$, suggesting that the effect size is not zero, which is a counter-intuitive conclusion to draw simultaneously with the conclusion of equivalence.

Note that it is possible to calculate a Bayes factor for the same band around $\delta = 0$ of $2c$, but there is no need as the evidence in favor of $\delta = 0$ can be quantified directly. Because of this, the Bayes factor approach simplifies testing for equivalence, such that no arbitrary band needs to be established. Furthermore, one is allowed to make claims about the absence of an effect, something that is not possible with the conventional NHST approach. We will illustrate this approach with a reanalysis of data reported in [36].

Equivalence between short- and long-term storage of red-cells on the Multiple Organ Dysfunction Score

In Steiner et al. [36], the authors examine the properties of the duration of storage for red-cells intended for transfusion. The authors assert that there is considerable uncertainty about potentially deleterious effects of long-term storage of red-cells before transfusion. In this study, the authors examine whether there are differences on the Multiple Organ Dysfunction Score (MODS) between patients that receive red-cells for transfusion that have been stored a short time (10 days or less) versus a long time (21 days or more). Although the authors do not explicitly conduct an equivalence design, the implicit goal seems to be to test whether or not longer storage of red cells is harmful. The authors conclude that "duration of red-cell storage was not associated with significant differences in the change in MODS". The authors support this claim with a p-value of 0.44.

The application of the conventional NHST does not allow us to make any definite claims about the absence of a difference. In this demonstration, we reanalyze these data and calculate a Bayes factor to quantify the strength of evidence for equivalence provided by the data. For the analyses, we make use of the data presented in Table 2 of the manuscript. In this table, the means are rounded to one decimal. To approximate the original analysis as accurately as possible, we work with means of 8.516 and 8.683 (reported means are 8.5 and 8.7, respectively) to approximate the reported p-value as closely as possible. For calculation of the Bayes factor, we assume a Cauchy prior centered on $\delta = 0$.

The reanalysis for the equivalence test of short- versus long-term storage of red-cells now proceeds as follows:

- Calculate the t-statistic for the null-hypothesis that the difference in MODS scores between patients that were administered red-cells that were stored short- versus long is zero: $t = \frac{8.516-8.683}{3.6\sqrt{1/538+1/560}} = -0.77$.
- We use Equation 4 to calculate a two-sided Bayes factor quantifying the relative likelihood of the hypotheses $d = 0$ versus $d \neq 0$ given the data (BF_{01}). This leads to $BF_{01} = 11.04$, indicating that the null-hypothesis is over 11 times more likely than the two-sided alternative, given the data.

These and other equivalence Bayes factors can be obtained by providing values for the sample size n, group means, and group sds to the script: $n_1 = 538$, $n_2 = 560$, $M_1 = 8.516$, and $M_2 = 8.683$, $sd_1 = 3.6$, and $sd_2 = 3.6$ (further details can be found in the annotated code). This reanalysis corroborates the finding of the original authors, but allows us to go beyond the original claim by stating we have found evidence in favor of equivalence between short-term and long-term storage of

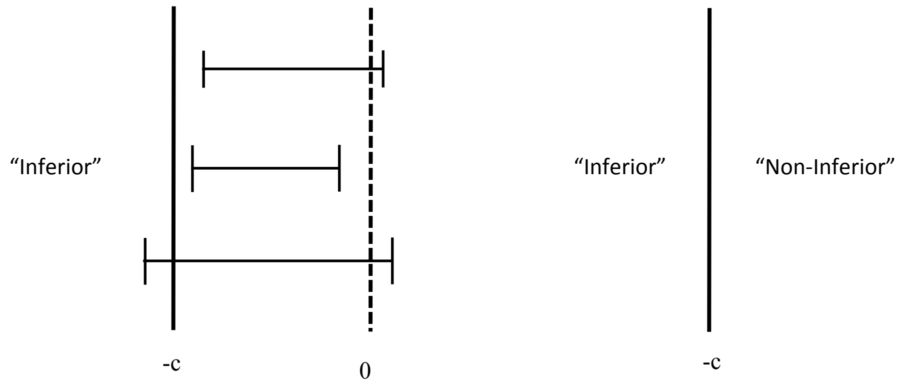

Figure 2 Non-inferiority design within an NHST framework (left) and a Bayesian framework (right). See text for details.

red-cells as far as MODS scores are concerned. The Bayes factor lies between 3 and 20, which may be interpreted as positive evidence in favor of equivalence.

Steiner et al. [36] do not provide an equivalence margin, but it is important to stress that if they had, a Bayes factor for the relative likelihood of the population parameter being inside versus outside of this equivalence band can easily be calculated as well. Say, for instance, that $c = 0.05$, then the two-sided Bayes factor quantifying the relative likelihood of the nul hypothesis $-c < d < c$ versus the alternative hypothesis $d < -c$ or $d > c$ given the data is 19.09.

Bayes factors for non-inferiority designs

In a traditional NHST approach, a non-inferiority design specifies a null-hypothesis of $\delta = -c$, and requires a one-sided z- or t-test (or construction of a confidence interval). The crucial test is whether this test rejects or fails to reject inferiority (see Figure 2, left panel).

The NHST approach for non-inferiority houses some unfortunate inconsistencies. Take for instance the top confidence interval in the left panel. This is an example of a situation where the null-hypothesis of inferiority gets rejected. From an NHST perspective, there is nothing wrong with this conclusion, as the confidence interval overlaps with zero, making the conclusion "non-inferior" warranted within that framework.

By contrast, examine the middle confidence interval in the left panel. Again, the null-hypothesis of inferiority gets rejected. This time, the implications are a bit less clear, because the confidence interval does not overlap with zero. From an NHST perspective, one would simultaneously reject the inferiority hypothesis and a classical one-sided null-hypothesis, reaching opposite conclusions. This makes it somewhat unclear if the conclusion "non-inferior" is really warranted here.

Finally, the bottom confidence interval in the left panel shows the scenario where the "inferior" null-hypothesis cannot be rejected. From an NHST perspective, we are unable to draw any further conclusions: is the drug/treatment inferior, or was the trial underpowered?

The Bayesian approach is not hampered by these pitfalls in interpretation. A Bayesian is concerned with the two hypotheses depicted in the right panel of Figure 2. In [37] Bayesian approaches for non-inferiority trials are discussed (see also [38, 39, 40]), but discussion of the implementation of Bayes factors is limited to dichotomous data [41]. Here, we propose to calculate Bayes factors for continuous data, using the same principle as for superiority and equivalence designs illustrated above. The Bayes factor in this case quantifies the relative likelihood of the data having occurred given inferiority versus the likelihood of the data having occurred given non-inferiority.

Analogous to the Bayes factor for superiority/equivalence designs, we use the Cauchy prior distribution for effect size δ, centered on zero. The classical z- or t-statistics were evaluated against $\delta = -c$. In order to maintain the theoretical property of the prior being centered on zero as specified in Jeffreys work, we shift the center of the Cauchy prior distribution to c. The easiest way to see why this is so is by imagining adding c to all data-points, all hypotheses, and all distributions, so that we evaluate the t-test for null hypothesis $\delta = 0$. The resulting test statistic will not change, as the data and the hypotheses have shifted by the same amount, but the prior distribution is now centered at c. Equation 4 allows for different specifications (for instance, a prior centered on the non-inferiority margin), but for these examples, we will keep the prior consistent across the design types. We will illustrate this approach with two examples. We first reanalyze dichotomous data published in [42], and then reanalyze continuous data published in [43].

Non-inferiority of beta-lactam

In [42], the authors examine antibiotic treatments for patients with clinically suspected community-acquired pneumonia (CAP). Specifically, guidelines recommend supplementing administration of beta-lactam with either macrolides or fluoroquinolones. The authors state that there is limited evidence that macrolides and/or fluoroquinolones have added benefits over the administration of just beta-lactam. In this study, the authors "tested the noninferiority of the beta-lactam strategy to the beta-lactam-macrolide and fluoroquinolone strategies with respect to 90-day mortality using a noninferiority margin of 3 percentage points and a two-sided 90% confidence interval." The authors conclude that "the risk of death was higher by 1.9 percentage points (90% confidence interval [CI], -0.6 to 4.4) with the beta-lactam-macrolide strategy than with the beta-lactam strategy and lower by 0.6 percentage points (90% CI, -2.8 to 1.9) with the fluoroquinolone strategy than with the beta-lactam strategy. These results indicated noninferiority of the beta-lactam strategy."

Sample sizes in the beta-lactam, beta-lactam-macrolide, and beta-lactam-fluoroquinolone groups are 656, 739, and 888, respectively. The crude 90-day mortality was 9.0% (59 patients), 11.1% (82 patients), and 8.8% (78 patients), respectively, during these strategy periods. In this demonstration, we reanalyze these data and do two Bayesian tests for non-inferiority. For the analyses, we make the following assumptions:

- The critical non-inferiority tests compare two proportions. Like the original authors, we use the normal approximation for the sampling distribution of proportions. In all three groups, sample sizes are sufficiently large to make this a safe assumption.

- The Bayes factor approach requires specifying the non-inferiority margin in terms of effect size Cohen's d. Cohen's h for proportions has similar properties to Cohen's d for continuous data. Converting the 3 percentage points yields a Cohen's h of $2 * \arcsin(\sqrt{\frac{59+82}{656+739}}) - 2 * \arcsin(\sqrt{\frac{59+82}{656+739} - .03}) = 0.11$. Going forward, we will refer to this value as h.
- The equation we use to calculate the relevant Bayes factors, Equation 4, assumes a t-test statistic. For these sample sizes, the t-statistic is virtually indistinguishable from the Z-statistic provided by the normal approximation.

With these assumptions in place, the reanalysis for the beta-lactam versus beta-lactam-macrolide groups now proceeds as follows:

- Calculate the Z-statistic for the null-hypothesis that the difference in proportions of mortality in the beta-lactam group and the beta-lactam-macrolide group is .03: $Z = \frac{59/656 - 82/739 - .03}{\sqrt{(59+82)/(656+739) \times (1-(59+82)/(656+739)) \times (1/656+1/739)}} = -3.16$.
- We use Equation 4 to calculate a one-sided Bayes factor quantifying the relative likelihood of the hypotheses $h < 0.11$ versus $h = 0.11$ given the data (BF_{-h}), and to calculate a one-sided Bayes factor quantifying the relative likelihood of the hypotheses $h = 0.11$ versus $h > 0.11$ given the data (BF_{h+}).
- Finally, we use the principal of transitivity, $BF_{-+} = BF_{-h} \times BF_{h+}$. BF_{-+} quantifies the relative evidence for non-inferiority (difference in mortality rate is lower than 3 percentage points) versus inferiority (difference in mortality rate is higher than 3 percentage points), given the data. For these data, $BF_{-+} = 1307.76$, indicating that the non-inferiority hypothesis is over 1,300 times more likely than the inferiority hypothesis, given the data.

These and other non-inferiority Bayes factors for proportions can be obtained by providing values for the sample size n, mortality count k, and the non-inferiority margin to the script: $n_1 = 656$, $n_2 = 739$, $k_1 = 59$, $k_2 = 82$, and $NI_{mar} = 0.03$ (further details can be found in the annotated code). A similar reanalysis for the beta-lactam versus beta-lactam-fluoroquinolone groups yields $BF_{-+} = 39.07$, indicating that the non-inferiority hypothesis is almost 40 times more likely than the inferiority hypothesis, given the data. Thus, our results corroborate those of the original authors, we find non-inferiority for beta-lactam versus beta-lactam-macrolide and beta-lactam-fluoroquinolone. The Bayes factors allow us to make claims about the strength of evidence, with support for non-inferiority of beta-lactam compared to beta-lactam-fluoroquinolone being strong, and support for non-inferiority of beta-lactam compared to beta-lactam-macrolide being overwhelming.

The above example demonstrates calculation of the Bayes factor for non-inferiority trials with dichotomous outcome measures. We now turn to a second example of our approach that showcases the application of our method for outcome data that is measured on a continuous scale.

Non-inferiority of internet-delivered cognitive behavior therapy

In [43], the authors examine the efficacy of internet-delivered cognitive behavior therapy (ICBT) in the treatment of mild to moderate depression symptoms, specifically by comparing its effectiveness to the 'regular' group-based cognitive behavior therapy (CBT). Depression symptoms are measured with the self-rated version of the Montgomery-Asberg Depression Rating Scale (MADRS). The authors define inferiority as a two-point difference on the MADRS between CBT and ICBT. The authors assess non-inferiority directly post-treatment and in a three-year follow-up and conclude that "Results on the self-rated version of the Montgomery-Asberg Depression Scale showed significant improvements in both groups across time indicating non-inferiority of guided ICBT."

Sample sizes in the ICBT and CBT groups are 32 and 33 respectively post-treatment and 32 and 30 respectively in the three year follow-up. In this demonstration, we reanalyze these data and do two Bayesian tests for non-inferiority. For the analyses, we make use of the data presented in Table 2 of the manuscript.

The reanalysis for the ICBT versus CBT groups now proceeds as follows:

- The Bayes factor approach requires specifying the non-inferiority margin in terms of effect size Cohen's d. Converting the 2 point difference yields a Cohen's d of $d_{post} = 2/\sqrt{\frac{31*9.8^2+32*8^2}{63}} \approx 0.22$ for the post-treatment group and $d_3 = 2/\sqrt{\frac{31*7.6^2+29*8.7^2}{60}} \approx 0.25$ for the three year follow-up group.
- Calculate the t-statistic for the null-hypothesis that the difference in MADRS scores in the ICBT group and the CBT groups is 2: $t_{post} = \frac{13.6-17.1-2}{\sqrt{\frac{31*9.8^2+32*8^2}{63}} \times \sqrt{1/32+1/33}} = -2.48$.
- We use Equation 4 to calculate a one-sided Bayes factor quantifying the relative likelihood of the hypotheses $d_{post} < 0.22$ versus $d_{post} = 0.22$ given the data (BF_{-d}), and to calculate a one-sided Bayes factor quantifying the relative likelihood of the hypotheses $d_{post} = 0.22$ versus $d_{post} > 0.22$ given the data (BF_{d+}).
- Finally, we use the principal of transitivity, $BF_{-+} = BF_{-d} \times BF_{d+}$. BF_{-+} quantifies the relative evidence for non-inferiority (difference in depression scores is lower than 2 points) versus inferiority (difference in depression scores is higher than 2 points), given the data. For these data, $BF_{-+} = 90.52$, indicating that the non-inferiority hypothesis is over 90 times more likely than the inferiority hypothesis, given the data.

These and other non-inferiority Bayes factors for continuous data can be obtained by providing values for the sample size n, group means, group sds, and the non-inferiority margin to the script: $n_1 = 32$, $n_2 = 33$, $M_1 = 13.6$, and $M_2 = 17.1$, $sd_1 = 9.8$, $sd_2 = 8$, and $NI_{mar} = 2$ (further details can be found in the annotated code). A similar reanalysis for the three year follow-up non-inferiority test yields $BF_{-+} = 353.61$. Thus, our results corroborate those of the original authors, we find non-inferiority for ICBT versus CBT directly after treatment and in a three-year follow-up. Note that despite the relatively small sample size, the Bayes factors quantifying strength of evidence in favor of non-inferiority are substantial, highlighting one of the advantages of quantifying evidence with Bayes factors: a clear measure of the strength of evidence for one hypothesis relative to another that can be used to compare evidence across studies.

Conclusions

In this paper, we showed worked examples of the application of default Bayes factors to superiority, non-inferiority, and equivalence designs. In each of these cases, we believe that application of Bayes factors brings significant advantages. For superiority and equivalence designs alike, it is possible to explicitly quantify evidence in favor of the null hypothesis. For equivalence studies, specification of a potentially arbitrary band of equivalence is no longer necessary. For non-inferiority and equivalence designs alike, the interpretational hazard of simultaneously claiming non-inferiority/equivalence on one hand, but rejecting the null hypothesis of an effect size of zero on the other hand, disappears. The Bayes factor offers a way to quantify each of these types of evidence in a compelling and straightforward way.

Some caveats to this kind of analysis should be considered. First of all, much like in NHST, one-tailed and two-tailed tests can give strikingly different results when the difference between groups is in the opposite direction of the one specified by the one-tailed test. We have seen such a scenario in our example for the superiority design, where we obtained $BF_{0-} = 31.48$ indicating that the null hypothesis is over 30 times more likely than superiority for our one-tailed test and $BF_{10} = 2.24$ indicating that the alternative hypothesis is slightly more likely than the null hypothesis for our two-tailed test. An illustration of how such discrepancies come about is given in Figure 3.

The two panels demonstrate a two-sided Bayes factor (left) and a one-sided Bayes factor (right), calculated based on the same hypothetical prior ($N(0, 1)$) and posterior ($N(1, 0.75)$) distributions for effect size. In both cases, the Bayes factor is obtained by dividing the density of the prior, evaluated at zero, by the density of the posterior, evaluated at zero (i.e., the black dot divided by the red dot). For the one-sided Bayes factor, both distributions are truncated at zero. Because both distributions are normalized to have a density of 1, the effect of this truncation is especially strong for a distribution that falls almost entirely inside the truncated region, such as in the posterior distribution of our example data here and in the [35] superiority design data we reanalyzed. As a result, the Bayes factors in Figure 3 lead to opposite conclusions, depending on whether the test was designed to be one-tailed or two-tailed. This example demonstrates that it is crucial to think about the hypotheses one wishes to test and the direction of testing before one obtains the data. Similar considerations apply when testing within the classical NHST framework.

Secondly, in NHST the status of $\alpha = 0.05$ is well established as a cut-off for significance (but see citeBenjaminEtAl2018). Bayesian inference does not have such universally agreed upon decision thresholds. Although different suggestions are offered in the literature [27, 18, 44], the authors caution against too rigid interpretation of these labels. We would argue that every cut-off value one chooses is to some extent arbitrary. With Bayes factors, one can at least choose a symmetrical cut-off score (for instance, we test until one hypothesis is 20 times more likely than the other given the data, so $BF_{10} = 20$ or $BF_{10} = 1/20 = 0.05$), whereas no such symmetry can be obtained with a p-value.

Thirdly, there are different ways to calculate Bayes factors [41]. Arguably the most important determinant for differences in Bayes factors stem from the choice of the underlying prior. Taking as an example the category of Bayes factors that assume a prior distribution on effect size, a prior that places a relatively high weight on an effect size of zero (i.e., is tightly peaked around zero), will lead to a relatively

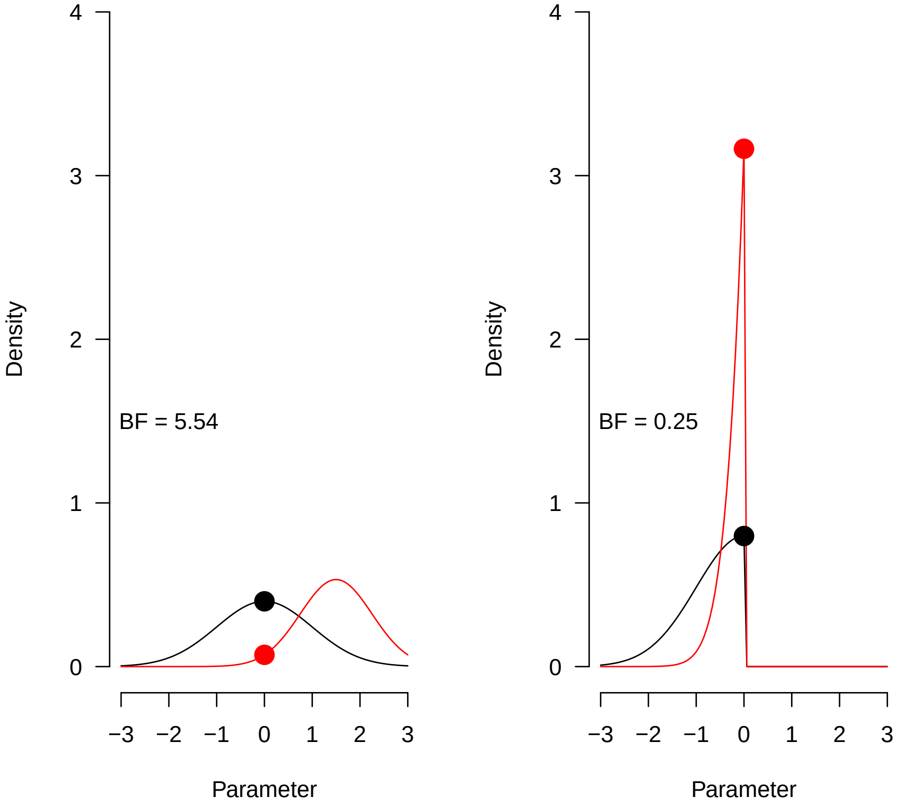

Figure 3 Hypothetical two-sided Bayes factor (left) and one-sided Bayes factor (right) for the same prior (black) and posterior (red) distribution on effect size.

large Bayes factor in favor of the alternative hypothesis if the sample effect size is relatively different from zero. For reasonable priors, the effect of the choice of prior on the Bayes factor appears to be mostly quantitative and unlikely to alter the qualitative conclusions [28]. Nevertheless, in specific applications, these default prior analyses can be supplemented by substantive knowledge based on earlier experience. With a more informative prior distribution, the alternative hypothesis will make different predictions, and a comparison with the null hypothesis will therefore yield a different Bayes factor. The more informed the prior distribution, the more specific the model predictions, and the more risk the analyst is willing to take. Highly informed prior distributions need to be used with care, as they may exert a dominant effect on the posterior distribution, making it difficult to "recover" once the data suggest that the prior was ill-conceived. With informed prior distributions, it is wise to perform a robustness analysis to examine the extent to which different modeling choices lead to qualitatively different outcomes.

Despite these considerations, our paper offers an easy way of calculating Bayes factors for superiority, equivalence, and non-inferiority designs that is consistent across methods and scale of the outcome measure. With increasing accessibility of software aimed to conduct Bayesian inference [33], the absence of tools necessary to obtain Bayes factors is no longer a reason to refrain from using Bayesian analyses. We recommend standard consideration of Bayesian inference in clinical trials for obtaining strength of evidence that is consistent across studies.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

DvR drafted the original manuscript and conducted the formal analyses. DvR and JPAI conceptualized the project. JT, RM, and JPAI reviewed and edited the manuscript.

Availability of data and materials

Code for reproducing the analyses reported in the manuscript may be obtained from `https://osf.io/8br5g/`.

Funding

This research was supported by a Dutch scientific organization VIDI grant to DvR (016.Vidi.188.001).

Author details

[1]University of Groningen, Department of Psychology, Grote Kruisstraat 2/1, Heymans Building, 9712 TS Groningen, The Netherlands. [2]University Medical Center Groningen, Groningen, The Netherlands. [3]Departments of Medicine, of Health Research and Policy, of Biomedical Data Science, and of Statistics, and Meta-Research Innovation Center, Stanford, United States.

References

1. Senn, S.S.: Statistical Issues in Drug Development, 2nd edn. John Wiley & Sons, Hoboken (NJ) (2008)
2. Piaggio, G., Elbourne, D.R., Pocock, S.J., Evans, S.J.W., Altman, D.G., Consort Group: Reporting of noninferiority and equivalence randomized trials: Extension of the CONSORT 2010 statement. JAMA **308**, 2594–2604 (2012)
3. Chamberlain, D.B., Chamberlain, J.M.: Making Sense of a Negative Clinical Trial Result: A Bayesian Analysis of a Clinical Trial of Lorazepam and Diazepam for Pediatric Status Epilepticus. Annals of Emergency Medicine **69**, 117–124 (2017)
4. Greene, W.L., Concato, J., Feinstein, A.R.: Claims of equivalence in medical research: are they supported by the evidence? Annals of Internal Medicine **132**, 715–722 (2000)
5. Macleod, M.R., Michie, S., Roberts, I., Dirnagl, U., Chalmers, I., Ioannidis, J.P.A., Salman, R.A.-S., Chan, A.-W., Glasziou, P.: Biomedical research: Increasing value, reducing waste. The Lancet **383**, 101–104 (2014)
6. Mueller, P.S., Montori, V.M., Bassler, D., Koenig, B.A., Guyatt, G.H.: Ethical issues in stopping randomized trials early because of apparent benefit. Annals of Internal Medicine **146**, 878–881 (2007)
7. Berry, D.A.: Bayesian clinical trials. Nature Reviews Drug Discovery **5**, 27–36 (2006)
8. Zohar, S., Latouche, A., Taconnet, M., Chevret, S.: Software to compute and conduct sequential Bayesian phase I or II dose-ranging clinical trials with stopping rules. Computer methods and programs in biomedicine **72**, 117–125 (2003)
9. Kalil, A.C., Sun, J.: Low-dose steroids for septic shock and severe sepsis: The use of Bayesian statistics to resolve clinical trial controversies. Intensive Care Medicine **37**, 420–429 (2011)
10. Lee, J.J., Chu, C.T.: Bayesian clinical trials in action. Statistics in Medicine **31**, 2955–2972 (2012)
11. van Ravenzwaaij, D., Ioannidis, J.P.A.: A simulation study of the strength of evidence in the recommendation of medications based on two trials with statistically significant results. PLoS One **12**, 0173184 (2017)
12. Monden, R., Roest, A.D., van Ravenzwaaij, D., Wagenmakers, E.-J., Morey, R., Wardenaar, K.J., de Jonge, P.: The comparative evidence basis for the efficacy of second-generation antidepressants in the treatment of depression in the US: A Bayesian meta-analysis of Food and Drug Administration reviews. Journal of Affective Disorders **2018**, 393–398 (2018)
13. Food and Drug Administration: Guidance for Industry and FDA Staff: Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials. U.S. Department of Health and Human Services, Washington, DC (2010). https://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071121.pdf
14. Yin, G., Lam, C.K., Shi, H.: Bayesian randomized clinical trials: From fixed to adaptive design. Contemporary Clinical Trials (in press)
15. Luthra, S.: The scientific foundation, rationale and argument for a nonfrequentist bayesian analysis in clinical trials in coronary artery disease. Heart, Lung and Circulation **24**, 614–616 (2015)
16. Zaslavsky, B.G.: Bayesian hypothesis testing in two-arm trials with dichotomous outcomes. Biometrics **69**, 157–163 (2013)
17. Sanders, G.D., Inoue, L., Samsa, G., Kulasingam, S., Matchar, D.: Use of Bayesian Techniques in Randomized Clinical Trials: A CMS Case Study. Agency for Healthcare Research and Quality (US), Rockville, MD (2009). http://www.ncbi.nlm.nih.gov/books/NBK253213/
18. Jeffreys, H.: Theory of Probability. Oxford University Press, Oxford, UK (1961)
19. Goodman, S.N.: Toward evidence-based medical statistics. 2: The Bayes factor. Annals of Internal Medicine **130**, 1005–1013 (1999)

20. Rouder, J.N.: Optional stopping: No problem for Bayesians. Psychonomic Bulletin & Review **21**, 301–308 (2014)
21. Edwards, W., Lindman, H., Savage, L.J.: Bayesian statistical inference for psychological research. Psychological Review **70**, 193–242 (1963)
22. O'Hagan, A.: Dicing with the unknown. Significance **1**, 132–133 (2004)
23. Pezeshk, H.: Bayesian techniques for sample size determination in clinical trials: A short review. Statistical Methods in Medical Research **12**, 489–504 (2003)
24. Rouder, J.N., Speckman, P.L., Sun, D., Morey, R.D., Iverson, G.: Bayesian t–tests for accepting and rejecting the null hypothesis. Psychonomic Bulletin & Review **16**, 225–237 (2009)
25. O'Hagan, A., Forster, J.: Kendall's Advanced Theory of Statistics Vol. 2B: Bayesian Inference (2nd ed.). Arnold, London (2004)
26. Berger, J.O.: Bayes factors. In: Kotz, S., Balakrishnan, N., Read, C., Vidakovic, B., Johnson, N.L. (eds.) Encyclopedia of Statistical Sciences, Vol. 1 (2nd Ed.), pp. 378–386. Wiley, Hoboken, NJ (2006)
27. Kass, R.E., Raftery, A.E.: Bayes factors. Journal of the American Statistical Association **90**, 773–795 (1995)
28. Gronau, Q.F., Ly, A., Wagenmakers, E.-J.: Informed Bayesian T-tests. Manuscript submitted for publication (2018)
29. Gönen, M., Johnson, W.O., Lu, Y., Westfall, P.H.: The Bayesian two-sample t test. The American Statistician **59**, 252–257 (2005)
30. Liang, F., Paulo, R., Molina, G., Clyde, M.A., Berger, J.O.: Mixtures of g priors for bayesian variable selection. Journal of the American Statistical Association **103**, 410–423 (2008)
31. Morey, R.D., Rouder, J.N., Jamil, T., Urbanek, S., Forner, K., Ly, A.: Bayes Factor (Version 0.9.12-4.1) [computer software] (2018). `https://CRAN.R-project.org/package=BayesFactor`
32. Gu, X., Mulder, J., Hoijtink, H.: Approximate adjusted fractional Bayes factors: A general method for testing informative hypotheses. British Journal of Mathematical and Statistical Psychology (in press)
33. The JASP Team: JASP (Version 0.8.6) [computer software] (2018). `https://jasp-stats.org/`
34. Hoekstra, R., Monden, R., van Ravenzwaaij, D., Wagenmakers, E.-J.: Bayesian reanalysis of null results reported in the New England Journal of Medicine: Strong yet variable evidence for the absence of treatment effects. PLoS One **13**, 0195474 (2018)
35. Skjerven, H.O., Hunderi, J.O.G., Brügmann-Pieper, S.K., Brun, A.C., Engen, H., Eskedal, L., Haavaldsen, M., Kvenshagen, B., Lunde, J., Rolfsjord, L.B., Carlsen, K.C.L.: Racemic adrenaline and inhalation strategies in acute bronchiolitis. New England Journal of Medicine **368**, 2286–2293 (2013)
36. Steiner, M.E., Ness, P.M., Assmann, S.F., Triulzi, D.J., Sloan, S.R., Delaney, M., Granger, S., Bennett-Guerrero, E., Blajchman, M.A., Scavo, V., Stowell, C.P.: Effects of red-cell storage duration on patients undergoing cardiac surgery. New England Journal of Medicine **372**, 1419–1429 (2015)
37. Gamalo-Siebers, M., Gao, A., Lakshminarayanan, M., Liu, G., Natanegara, F., Railkar, R., Schmidli, H., Song, G.: Bayesian methods for the design and analysis of noninferiority trials. Journal of Biopharmaceutical Statistics **26**, 823–841 (2016)
38. Gamalo, M.A., Wu, R., Tiwari, R.C.: Bayesian approach to non-inferiority trials for normal means. Statistical Methods in Medical Research **25**, 221–240 (2016)
39. Ghosh, P., Nathoo, F., Gönen, M., Tiwari, R.C.: Assessing noninferiority in a three-arm trial using the Bayesian approach. Statistics in Medicine **30**, 1795–1808 (2011)
40. Daimon, T.: Bayesian sample size calculations for a non-inferiority test of two proportions in clinical trials. Contemporary Clinical Trials **29**, 507–516 (2008)
41. Osman, M., Ghosh, S.K.: Semiparametric Bayesian testing procedure for noninferiority trials with binary endpoints. Journal of Biopharmaceutical Statistics **21**, 920–937 (2011)
42. Postma, D.F., Van Werkhoven, C.H., Van Elden, L.J.R., Thijsen, S.F.T., Hoepelman, A.I.M., Kluytmans, J.A.J.W., Boersma, W.G., Compaijen, C.J., Van Der Wall, E., Prins, J.M., Oosterheert, J.J., Bonten, M.J.M., the CAP-START Study Group: Antibiotic treatment strategies for community-acquired pneumonia in adults. New England Journal of Medicine **372**, 1312–1323 (2015)
43. Andersson, G., Hesser, H., Veilord, A., Svedling, L., Andersson, F., Sleman, O., Mauritzson, L., Sarkohi, A., Claesson, E., Zetterqvist, V., Lamminen, M., Eriksson, T., Carlbring, P.: Randomised controlled non-inferiority trial with 3-year follow-up of internet-delivered versus face-to-face group cognitive behavioural therapy for depression. Journal of Affective Disorders **151**, 986–994 (2013)
44. Weiss, R.: Bayesian sample size calculations for hypothesis testing. The Statistician **46**, 185–191 (1997)