A Comparison of the Bayesian and Frequentist Approaches to Equivalence, Non-Inferiority, and Superiority Designs

Maximilian Linde

(s2512491)

Master (major) thesis

MSc Program of Behavioural and Cognitive Neurosciences

(C-track; Cognitive Neuroscience and Cognitive Modelling)

Faculty of Science and Engineering

University of Groningen

July, 2019

Supervisor: Dr. Don van Ravenzwaaij

Second evaluator: Dr. Jorge N. Tendeiro

Abstract

Clinical trials often seek to determine the equivalence, non-inferiority, or superiority of an experimental condition (e.g., a new drug) compared to a control condition (e.g., a placebo or an already existing drug). The use of frequentist statistical methods to analyse data for these types of designs is widespread. Importantly, however, frequentist inference has several limitations. Bayesian statistics remedies these shortcomings and allows for intuitive interpretations. The goal of the present article is twofold. First, we present baymedr, an R package that provides user-friendly tools for the computation of Bayes factors for equivalence, non-inferiority, and superiority designs (see also van Ravenzwaaij, Monden, Tendeiro, & Ioannidis, 2019). Second, we conducted simulations to contrast the performances of the Bayesian and frequentist equivalence and non-inferiority tests. We generated data sets with various true population effect sizes and sample sizes, which were analysed by the frequentist and Bayesian equivalence and non-inferiority tests. The resulting p-values and Bayes factors formed the basis for subsequent receiver operating characteristic (ROC) analyses. The classification performances of the Bayesian tests were higher compared to their frequentist counterparts. Generally, the frequentist equivalence and non-inferiority tests demanded a high sample size to reach a proper power. Together with the theoretical advantages of Bayesian inference, this leads us to propose the adaptation of our state-of-the-art Bayesian tools for the analysis of equivalence, non-inferiority, and superiority studies.

Keywords: Bayes factor, baymedr, equivalence, non-inferiority, superiority

A Comparison of the Bayesian and Frequentist Approaches to Equivalence, Non-Inferiority, and Superiority Designs

Introduction

Researchers generally agree that the clinical trial is the best method to compare the effects of medications and treatments (e.g., E. Christensen, 2007; Friedman, Furberg, DeMets, Reboussin, & Granger, 2010). Although clinical trials are similar in design, there are diverse research questions that shall be answered. Commonly, clinical trials seek to determine the superiority, equivalence, or non-inferiority of an experimental condition (e.g., a new medication) compared to a control condition (e.g., a placebo or an already existing medication [active control]; Lesaffre, 2008; Piaggio, Elbourne, Pocock, Evans, & Altman, 2012). Usually, the frequentist approach to statistics forms the framework in which data for these research designs are analysed (Chavalarias, Wallach, Li, & Ioannidis, 2016). In particular, a null hypothesis significance test (NHST) is conducted and evidence is quantified by a p-value. The resulting p-value represents the probability of obtaining a test statistic at least as extreme as the one observed, assuming that the null hypothesis is true. In other words, the p-value is an indicator of the unusualness of the obtained test statistic under the null hypothesis, forming a 'proof by contradiction' (R. Christensen, 2005). If the p-value is smaller than the predefined Type I error rate (α), which is standardly set to $\alpha = .05$ (but see, e.g., Benjamin et al., 2018), we reject the null hypothesis; otherwise we fail to reject the null hypothesis.

Importantly, however, the NHST approach to inference has been increasingly critiqued on several grounds due to certain limitations and erroneous interpretations of p-values (e.g., Berger & Sellke, 1987; Cohen, 1994; Dienes, 2011; Gigerenzer, Krauss, & Vitouch, 2004; Goodman, 1999a, 1999b, 2008; Loftus, 1996; van Ravenzwaaij & Ioannidis, 2017; Wagenmakers, 2007; Wagenmakers et al., 2018; Wasserstein & Lazar, 2016; Wetzels et al., 2011). As a result, many methodologists have argued that frequentist inference should be mostly abandoned. An alternative is the Bayesian approach to data analysis, which is a school of statistics that is based on the idea that

the credibility of well defined possibilities, parameters, or models (e.g., null and alternative hypotheses) are updated based on new observations (Kruschke, 2015). With exploding computational power and the rise of Markov chain Monte Carlo methods (e.g., Gilks, Richardson, & Spiegelhalter, 1995; van Ravenzwaaij, Cassey, & Brown, 2018), that can be utilised to estimate probability distributions that cannot be determined analytically, in the past few decades, applications of Bayesian inference became tractable. Indeed, Bayesian methods are gaining more and more traction in the biomedical field (Berry, 2006) and other disciplines (e.g., psychology; van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, & Depaoli, 2017).

Despite of this slowly developing paradigm shift away from frequentist towards Bayesian inference, a majority of biomedical research still employs frequentist statistical techniques (Chavalarias et al., 2016). To some extent, this might be due to a biased statistical education in favour of frequentist inference. Moreover, researchers might perceive the conduction of NHSTs and the reporting of p-values as prescriptive and, hence, adhere to this convention (see, e.g., Winkler, 2001). We believe that one of the most crucial factors is the unavailability of easy-to-use Bayesian tools and software, leaving Bayesian data analysis largely to statistical experts. Fortunately, important advances have been made towards user-friendly interfaces for Bayesian analyses with the release of the BayesFactor software (Morey & Rouder, 2018), written in R (R Core Team, 2019), and point-and-click software like JASP (JASP Team, 2018) and Jamovi (The jamovi project, 2019), both of which are based to some extent on the BayesFactor software (Morey & Rouder, 2018). However, these tools were mainly developed for research designs in the social sciences. Bayesian tools and corresponding accessible software for the analysis of common biomedical research designs (e.g., superiority, equivalence, and non-inferiority) are still missing and, thus, urgently needed.

In this article, we describe how to perform Bayesian inference for superiority, equivalence, and non-inferiority designs and compare the performances of the Bayesian equivalence and non-inferiority tests to the corresponding conventional frequentist counterparts by means of a set of simulations. In the first part of this article, we outline

the frequentist conceptualisation of superiority, equivalence, and non-inferiority designs. Subsequently, a discussion of the key disadvantages and potential pitfalls of this approach motivates our conviction that Bayesian inferential techniques are better suited for these research designs. We provide and introduce baymedr (available at https://github.com/maxlinde/baymedr), an open-source software written in R (R Core Team, 2019) for the computation of Bayes factors (e.g., Jeffreys, 1939, 1948, 1961; Kass & Raftery, 1995) for common biomedical designs, and explain its implementation and usage. In the second part, we contrast the performances of Bayesian and frequentist inferential techniques for equivalence and non-inferiority designs through a set of simulations. Their decision performances are compared in various situations through receiver operating characteristic curves. Note that we did not conduct simulations for the superiority design because similar simulations have already been done by van Ravenzwaaij and Ioannidis (2018).

Frequentist Inference for Superiority, Equivalence, and Non-Inferiority Designs

The superiority, equivalence, and non-inferiority tests are concerned with research settings in which two conditions (e.g., control and experimental) are compared on some outcome measure (E. Christensen, 2007; Lesaffre, 2008). For instance, researchers might want to investigate whether a new antidepressant medication is superior, equivalent, or non-inferior compared to a well-established antidepressant. The three research designs are quite similar in that they all contain some form of the t-test; however, they differ on the precise specification of the t-test. In the following, we will assume that higher scores on the measure of interest represent a more favourable outcome (i.e., superiority or non-inferiority) than lower scores.

The Superiority Design

The superiority design tests whether the experimental condition is superior to or better than the control condition. Conceptually, the superiority design consists of a one-sided test due to its inherent directionality. However, researchers often conduct a two-sided test instead and confirm afterwards that the results follow the expected

direction. Given a one-sided test, the null hypothesis states that the true population effect size is zero, whereas the alternative hypotheses states that the true population effect size is larger than zero:

$$\mathcal{H}_0 : \delta = 0$$
$$\mathcal{H}_1 : \delta > 0;$$

with a two-sided test, the null hypothesis is the same as in the one-sided test and the alternative hypothesis states that the true population effect size is unequal to zero:

$$\mathcal{H}_0 : \delta = 0$$
$$\mathcal{H}_1 : \delta \neq 0,$$

where δ represents the true population effect size between the experimental and the control conditions. To test these hypotheses, either a one- or two-tailed t-test is conducted. If the resulting p-value is smaller than α (and the effect size goes in the expected direction in case of the two-tailed test), we reject the null hypothesis and superiority of the experimental group is established.

The Equivalence Design

The equivalence design tests whether the experimental and control conditions are practically equivalent. Evidently, there are multiple approaches to equivalence testing (e.g., the power approach; see Meyners, 2012, for an accessible overview). A comprehensive treatment of all approaches is beyond the scope of this article. Here, we will focus on the two one-sided tests procedure (TOST; Lakens, 2017; Lakens, Scheel, & Isager, 2018; Schuirmann, 1987). An equivalence interval must be defined, which can be based, for example, on the smallest effect size of interest (SESOI; e.g., Lakens, 2017; Lakens et al., 2018). The specification of the equivalence interval is not a statistical question; thus, it should be set by experts in the respective fields (Meyners, 2012; Schuirmann, 1987) or comply with regulatory guidelines (Garrett, 2003). Importantly,

however, the equivalence interval should be determined prior to data analysis. As the name implies, TOST involves the conduction of two one-sided t-tests, each one with its own null and alternative hypotheses. For the first NHST, the null hypothesis states that the true population effect size is smaller than the lower boundary of the equivalence interval, whereas the alternative hypothesis states that the true population effect size is larger than the lower boundary of the equivalence interval; for the second NHST, the null hypothesis is that the true population effect size is larger than the upper boundary of the equivalence interval, whereas the alternative hypothesis is that the true population effect size is smaller than the upper boundary of the equivalence interval. These hypotheses can be summarised as follows:

$$\mathcal{H}_0 : \delta < \Delta_L \text{ OR } \quad \delta > \Delta_U$$
$$\mathcal{H}_1 : \delta > \Delta_L \text{ AND } \delta < \Delta_U,$$

where Δ_L and Δ_U represent the lower and upper boundaries of the equivalence interval, respectively. Two p-values (p_1 and p_2) result from the conduction of the two one-sided t-tests. We reject the null hypothesis of non-equivalence and, thus, establish equivalence if $\max(p_1,\ p_2) < \alpha$ (Meyners, 2012; Walker & Nowacki, 2011). In other words, both tests need to reach statistical significance.

The Non-Inferiority Design

In some situations, we just want to test that the experimental condition is non-inferior or not worse than the control condition by a certain amount. This is the goal of the non-inferiority design, which consists of a one-tailed test. Realistic applications might include a new medication that has fewer undesirable adverse effects (e.g., Chadwick, 1999), is cheaper (see, e.g., Kaul & Diamond, 2006, for a discussion), or is easier to administer than the old medication (e.g., Van de Werf et al., 1999). In these cases, we need to ponder the cost of a somewhat lower or equal effectiveness of the new treatment with the value of the just mentioned benefits (Hills, 2017). The null hypothesis states that the true population effect size is equal to a predetermined

threshold, whereas the alternative hypothesis states that the true population effect size is higher than this threshold:

$$\mathcal{H}_0 : \delta = -nim$$
$$\mathcal{H}_1 : \delta > -nim,$$

where *nim* represents the non-inferiority margin. As the equivalence interval, the non-inferiority margin should be defined a priori. The null hypothesis is rejected and non-inferiority established if the resulting p-value is smaller than α.

Limitations of Frequentist Inference

Tests of superiority, equivalence, and non-inferiority have great value in biomedical research, no doubt. It is the way researchers conduct their statistical analyses that, we argue, should be reconsidered. The use of NHSTs and p-values implicates several inevitable limitations in general, of which we will describe two, but also interpretational difficulties specific to superiority, equivalence, and non-inferiority designs.

First, researchers need to obey a predetermined sampling plan (e.g., Rouder, 2014; Schönbrodt & Wagenmakers, 2018; Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017). That is, it is not legitimate to examine the data during data collection and decide based on interim results to stop data collection (e.g., because the p-value is already lower than α) or to continue data collection beyond the predetermined sample size (e.g., because the p-value almost reaches statistical significance). In principle, researchers can correct for the fact that they inspected the data by reducing the required significance threshold through one of several techniques (Ranganathan, Pramesh, & Buyse, 2016). However, such correction methods are rarely applied.

Second, with the frequentist framework, it is impossible to quantify evidence in favour of the null hypothesis (e.g., Gallistel, 2009; van Ravenzwaaij et al., 2019; Wagenmakers, 2007; Wagenmakers et al., 2018). Oftentimes, the p-value is erroneously interpreted as a posterior probability, in the sense that it represents the probability for the truth of the null hypothesis (e.g., Berger & Sellke, 1987; Gelman, 2013; Goodman,

2008). However, a non-significant p-value does not only occur when the null hypothesis is in fact true but also when the alternative hypothesis is actually true, yet we did not have enough power to detect an effect (Bakan, 1966; van Ravenzwaaij et al., 2019). As Altman and Bland (1995) put it: "Absence of evidence is not evidence of absence". Still, a large proportion of biomedical studies falsely claim equivalence based on statistically non-significant t-tests (Greene, Concato, & Feinstein, 2000). Yet, estimating evidence in favour of the null hypothesis is essential for certain designs like the equivalence test (e.g., Blackwelder, 1982; van Ravenzwaaij, Monden, Tendeiro, & Ioannidis, 2019; see also Hoekstra, Monden, van Ravenzwaaij, & Wagenmakers, 2018).

The TOST procedure (Lakens, 2017; Lakens et al., 2018; Schuirmann, 1987) for equivalence testing provides a workaround for the problem that evidence for the null hypothesis cannot be quantified with frequentist techniques by defining an equivalence interval around $\delta = 0$ and conducting two NHSTs. Without this interval the TOST procedure would inevitably fail (see Meyners, 2012, for an explanation of why this is the case). As we will see, the Bayesian equivalence test does not have this restriction; it allows for the specification of an interval null hypothesis as well as a point null hypothesis.

Moreover, the frequentist non-inferiority and equivalence tests can result in ambivalent findings. Suppose that we conduct a non-inferiority test with a non-inferiority margin of nim. It is possible that the confidence interval of the difference between the experimental and control conditions lies between nim and 0. In that situation, the experimental group is non-inferior with respect to nim but inferior relative to 0. A similar ambiguity is possible in the TOST equivalence test. Equivalence is established when the confidence interval of the difference between the experimental and control conditions fully lies between the two equivalence boundaries (i.e., Δ_L and Δ_U). In the special case where this confidence interval does not overlap with 0, we would still conclude equivalence, although the confidence interval indicates the presence of an effect. Here again, we will see that the Bayesian approaches to non-inferiority and equivalence designs eliminate these interpretational ambiguities (see also van

Ravenzwaaij et al., 2019).

Bayesian Inference for Superiority, Equivalence, and Non-Inferiority Designs

The Bayesian framework forms a school of statistics that is based on the idea that the credibility of well-defined parameters and models is updated when new observations are obtained (Kruschke, 2015); it is a logically sound method to update our beliefs about parameters. Bayesian inference can be divided into parameter estimation (e.g., estimating a population correlation) and model comparison (e.g., comparing the credibility of the null and alternative hypotheses) procedures (e.g., Kruschke & Liddell, 2018b). Here we will focus on the latter approach, which is usually accomplished with Bayes factors (e.g., Jeffreys, 1939, 1948, 1961; Kass & Raftery, 1995). In our exposition of Bayes factors in general and specifically for superiority, equivalence, and non-inferiority designs, we mostly refrain from complex equations and derivations. Formulas are only provided when we think that they help to communicate the ideas and concepts. We refer readers interested in the mathematics of Bayes factors to other sources (e.g., Jeffreys, 1961; Kass & Raftery, 1995; Lee & Wagenmakers, 2013; O'Hagan & Forster, 2004; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010). The precise derivation of Bayes factors for superiority, equivalence, and non-inferiority designs in particular is treated by van Ravenzwaaij, Monden, Tendeiro, and Ioannidis (2019; see also Gronau, Ly, & Wagenmakers, 2019).

The Bayes Factor

Let us suppose that we have two models or hypotheses ($\mathcal{H}_0$ and $\mathcal{H}_1$) that we want to contrast. We have prior beliefs about the credibilities of $\mathcal{H}_0$ and $\mathcal{H}_1$, which are given by the prior probabilities $pr\left(\mathcal{H}_0\right)$ and $pr\left(\mathcal{H}_1\right) = 1 - pr\left(\mathcal{H}_0\right)$. Now, we collect some data D. After having seen the data, we have posterior beliefs about the probabilities that $\mathcal{H}_0$ and $\mathcal{H}_1$ are true, which are given by the posterior probabilities $pr\left(\mathcal{H}_0|D\right)$ and $pr\left(\mathcal{H}_1|D\right)$. In other words, we update our prior beliefs about the credibilities of $\mathcal{H}_0$ and $\mathcal{H}_1$ by incorporating what the data dictates we should believe and arrive at our posterior beliefs. This relation is expressed in Bayes' rule. For the calculation of the

posterior probability of $\mathcal{H}_0$, we have:

$$pr\left(\mathcal{H}_0|D\right) = \frac{pr\left(D|\mathcal{H}_0\right) \times pr\left(\mathcal{H}_0\right)}{\sum\limits_i pr\left(D|\mathcal{H}_i\right) \times pr\left(\mathcal{H}_i\right)}, \tag{1}$$

where $i \in \{0,\ 1\}$, $pr\left(\mathcal{H}_0\right)$ represents the prior probability of $\mathcal{H}_0$, $pr\left(D|\mathcal{H}_0\right)$ denotes the likelihood of the data under $\mathcal{H}_0$, $\sum_i pr\left(D|\mathcal{H}_i\right) \times pr\left(\mathcal{H}_i\right)$ is the marginal likelihood (also called evidence; see, e.g., Kruschke, 2015), and $pr\left(\mathcal{H}_0|D\right)$ is the posterior probability of $\mathcal{H}_0$. Similarly, we can use Bayes' rule to update our prior belief about $\mathcal{H}_1$ by incorporating the data to arrive at our posterior belief about $\mathcal{H}_1$:

$$pr\left(\mathcal{H}_1|D\right) = \frac{pr\left(D|\mathcal{H}_1\right) \times pr\left(\mathcal{H}_1\right)}{\sum\limits_i pr\left(D|\mathcal{H}_i\right) \times pr\left(\mathcal{H}_i\right)}. \tag{2}$$

The marginal likelihood serves as a normalisation constant, ensuring that the sum of the posterior probabilities is 1. Without it, however, the posterior is still proportional to the product of the likelihood and the prior. Therefore, for $\mathcal{H}_0$ we can also write:

$$pr\left(\mathcal{H}_0|D\right) \propto pr\left(D|\mathcal{H}_0\right) \times pr\left(\mathcal{H}_0\right); \tag{3}$$

and similarly for $\mathcal{H}_1$, we can write:

$$pr\left(\mathcal{H}_1|D\right) \propto pr\left(D|\mathcal{H}_1\right) \times pr\left(\mathcal{H}_1\right), \tag{4}$$

where $\propto$ means 'is proportional to'.

Rather than using posterior probabilities for each hypothesis, we can form an odds ratio of the posterior probabilities for $\mathcal{H}_0$ and $\mathcal{H}_1$. Dividing the posterior probability of $\mathcal{H}_0$ by the posterior probability of $\mathcal{H}_1$ yields:

$$\underbrace{\frac{pr\left(\mathcal{H}_0|D\right)}{pr\left(\mathcal{H}_1|D\right)}}_{\text{Posterior odds}} = \underbrace{\frac{pr\left(D|\mathcal{H}_0\right)}{pr\left(D|\mathcal{H}_1\right)}}_{\text{Bayes factor}} \times \underbrace{\frac{pr\left(\mathcal{H}_0\right)}{pr\left(\mathcal{H}_1\right)}}_{\text{Prior odds}}. \tag{5}$$

The quantity $pr\left(\mathcal{H}_0|D\right)/pr\left(\mathcal{H}_1|D\right)$ represents the posterior odds and the term

$pr\left(\mathcal{H}_0\right)/pr\left(\mathcal{H}_1\right)$ is called the prior odds. To get the posterior odds, we have to multiply the prior odds by $pr\left(D|\mathcal{H}_0\right)/pr\left(D|\mathcal{H}_1\right)$, a quantity known as the Bayes factor (Jeffreys, 1939, 1948, 1961; Kass & Raftery, 1995). The Bayes factor is the amount by which we would update our prior odds to reach the posterior odds, after taking into consideration the data. For example, if we had prior odds of 2 and the Bayes factor is 24, then the posterior odds would be 48. In the special case where the prior odds is 1, the Bayes factor is equal to the posterior odds.

A major advantage of the Bayes factor is its ease of interpretation. For example, if the Bayes factor (BF_{01}, denoting the fact that $\mathcal{H}_0$ is in the numerator and $\mathcal{H}_1$ in the denominator) equals 10, the data are ten times more likely to have occurred under $\mathcal{H}_0$ compared to $\mathcal{H}_1$. With $BF_{01} = 0.2$, we can say that the data are five times more likely under $\mathcal{H}_1$ compared to $\mathcal{H}_0$ because we can simply take the reciprocal to change the hypothesis towards which the Bayes factor shall quantify evidence: $BF_{10} = 1/BF_{01}$. What constitutes enough evidence is subjective and certainly depends on the context. Nevertheless, rules of thumb for evidence thresholds were proposed. For instance, Kass and Raftery (1995) labelled Bayes factors between 1 and 3 as 'not worth more than a bare mention', Bayes factors between 3 and 20 as 'positive', those between 20 and 150 as 'strong', and anything above 150 as 'very strong', with corresponding thresholds for the reciprocals of the Bayes factors. An alternative classification scheme was already proposed before by Jeffreys (1961) with thresholds at 3, 10, 30, and 100 and similar labels (cf. Lee & Wagenmakers, 2013, for updated labels).

Of course, we need to define $\mathcal{H}_0$ and $\mathcal{H}_1$. There are multiple ways to do this but for the purpose of this manuscript we will express the hypotheses using effect sizes (δ; Rouder et al., 2009). For instance, we could compare the null hypothesis $\mathcal{H}_0 : \delta = \delta_0$ to a two-sided alternative hypotheses ($\mathcal{H}_1 : \delta \neq \delta_0$) or to one of two one-sided alternative hypotheses ($\mathcal{H}_1 : \delta < \delta_0$ or $\mathcal{H}_1 : \delta > \delta_0$). Alternatively, we could compare an interval hypothesis for the null hypothesis ($\mathcal{H}_0 : \Delta_L < \delta < \Delta_U$) with an interval hypothesis for the alternative hypothesis ($\mathcal{H}_1 : \delta < \Delta_L$ OR $\delta > \Delta_U$).

In the most general case, we can calculate the Bayes factor (i.e., BF_{01}) through

division of the posterior odds by the prior odds:

$$BF_{01} = \frac{\left(\frac{pr\left(\mathcal{H}_0|D\right)}{pr\left(\mathcal{H}_1|D\right)}\right)}{\left(\frac{pr\left(\mathcal{H}_0\right)}{pr\left(\mathcal{H}_1\right)}\right)} = \frac{\left(\frac{pr\left(\mathcal{H}_0|D\right)}{pr\left(\mathcal{H}_0\right)}\right)}{\left(\frac{pr\left(\mathcal{H}_1|D\right)}{pr\left(\mathcal{H}_1\right)}\right)}. \tag{6}$$

Accordingly, we can also calculate BF_{10}:

$$BF_{10} = \frac{\left(\frac{pr\left(\mathcal{H}_1|D\right)}{pr\left(\mathcal{H}_0|D\right)}\right)}{\left(\frac{pr\left(\mathcal{H}_1\right)}{pr\left(\mathcal{H}_0\right)}\right)} = \frac{\left(\frac{pr\left(\mathcal{H}_1|D\right)}{pr\left(\mathcal{H}_1\right)}\right)}{\left(\frac{pr\left(\mathcal{H}_0|D\right)}{pr\left(\mathcal{H}_0\right)}\right)}. \tag{7}$$

Calculating Bayes factors this way often involves solving complex integrals (cf. Wagenmakers et al., 2010), which we will not discuss here. Fortunately, there is a computational shortcut for the specific but very common scenario where we have a point null hypothesis $\mathcal{H}_0$ and an alternative hypothesis $\mathcal{H}_1$ that is free to vary.

For the purpose of illustration, let us suppose that we have an interval null hypothesis $\mathcal{H}_0 : \Delta_L < \delta < \Delta_U$ with a corresponding alternative hypothesis $\mathcal{H}_1 : \delta < \Delta_L$ OR $\delta > \Delta_U$. From equation 6 we see that the Bayes factor (i.e., BF_{01}) is obtained by calculating $\left(pr\left(\mathcal{H}_0|D\right)/pr\left(\mathcal{H}_0\right)\right) / \left(pr\left(\mathcal{H}_1|D\right)/pr\left(\mathcal{H}_1\right)\right)$. Similarly, equation 7 shows that we can calculate BF_{10} with $\left(pr\left(\mathcal{H}_1|D\right)/pr\left(\mathcal{H}_1\right)\right) / \left(pr\left(\mathcal{H}_0|D\right)/pr\left(\mathcal{H}_0\right)\right)$. Crucially, however, as the null interval decreases, $pr\left(\mathcal{H}_0|D\right)/pr\left(\mathcal{H}_0\right)$ will dominate the calculation of the Bayes factor because $pr\left(\mathcal{H}_1|D\right)/pr\left(\mathcal{H}_1\right)$ approaches 1. Put more formally, the limit of the ratio of the posterior and the prior area under $\mathcal{H}_1$, as the width of the null interval approaches 0 (i.e., a point null hypothesis), is 1:

$$\lim_{(\Delta_U - \Delta_L) \to 0} \frac{pr\left(\mathcal{H}_1|D\right)}{pr\left(\mathcal{H}_1\right)} = 1. \tag{8}$$

Therefore, this term can safely be ignored in the case where we have a point null hypothesis ($\mathcal{H}_0 : \delta = \delta_0$) that is nested within an alternative hypothesis that is free to vary ($\mathcal{H}_1 : \delta \neq \delta_0$, or $\mathcal{H}_1 : \delta < \delta_0$, or $\mathcal{H}_1 : \delta > \delta_0$). With this constellation of hypotheses, we can simplify the calculation of the Bayes factor through the calculation of the

Savage-Dickey density ratio (Dickey & Lientz, 1970; Kass & Raftery, 1995; see also Wagenmakers et al., 2010, for an intuitive introduction). To obtain the Bayes factor (i.e., BF_{01}) we can simply divide the density of the posterior of $\mathcal{H}_1$ at the point null hypothesis by the density of the prior of $\mathcal{H}_1$ at the point null hypothesis. Thus, we have:

$$BF_{01} = \frac{pr\left(D|\mathcal{H}_0\right)}{pr\left(D|\mathcal{H}_1\right)} = \frac{pr\left(\delta = \delta_0|D,\, \mathcal{H}_1\right)}{pr\left(\delta = \delta_0|\mathcal{H}_1\right)}, \tag{9}$$

where δ_0 corresponds to the point hypothesis of $\mathcal{H}_0$ (e.g., $\delta_0 = 0$). Conversely, we can calculate BF_{10} by dividing the density of the prior of $\mathcal{H}_1$ at the point null hypothesis by the density of the posterior of $\mathcal{H}_1$ at the point null hypothesis:

$$BF_{10} = \frac{pr\left(D|\mathcal{H}_1\right)}{pr\left(D|\mathcal{H}_0\right)} = \frac{pr\left(\delta = \delta_0|\mathcal{H}_1\right)}{pr\left(\delta = \delta_0|D,\, \mathcal{H}_1\right)}. \tag{10}$$

Note that the integral of a probability density function must always be 1. With the abrupt cutoff at the point null hypothesis in case of a one-sided alternative hypothesis, the form of the prior and posterior probability density functions might look quite differently in comparison to the probability density functions in case a two-sided alternative hypothesis is employed.

Until this point in our exposition, we were quite vague about the exact form of the prior for the effect size under $\mathcal{H}_1$. A specification of the alternative hypothesis as, for example, $\mathcal{H}_1 : \delta \neq 0$ is all that is needed in the frequentist approach; however, the Bayesian approach requires a precise definition of this prior within $\mathcal{H}_1$. In principle, the prior within $\mathcal{H}_1$ can be defined as desired, satisfying the subjective needs of the researcher for a specific research setting. In fact, this is a fundamental part of Bayesian inference because various priors allow us to express a theory or prior beliefs (e.g., Morey, Romeijn, & Rouder, 2016; Vanpaemel, 2010). Most commonly, however, default or objective priors are employed that aim to increase the objectivity in specifying the prior of the effect size under $\mathcal{H}_1$ (e.g., Jeffreys, 1961; Rouder et al., 2009).

The Default Bayes Factor

In the situation where we have a point null hypothesis and an alternative hypothesis that involves a range of values, Jeffreys (1961) proposed to use a Cauchy prior with a scale parameter of $r = 1$ for the effect size under $\mathcal{H}_1$. This Cauchy distribution is equivalent to a Student's t distribution with one degree of freedom and resembles a standard Normal distribution, except that the Cauchy distribution has less mass at the centre but instead heavier tails (see Fig 1; see also Rouder et al., 2009). The scale parameter r defines the width of the Cauchy distribution; that is, half of the mass lies between $-r$ and r. Mathematically, the Cauchy distribution is equivalent to a Normal distribution with a mean μ_δ and a variance σ^2_δ, which follows an inverse Chi-square distribution with one degree of freedom, and in which σ^2_δ is integrated out (Liang, Paulo, Molina, Clyde, & Berger, 2008; Rouder et al., 2009).

Choosing a Cauchy prior with a location parameter of 0 and a scale parameter of

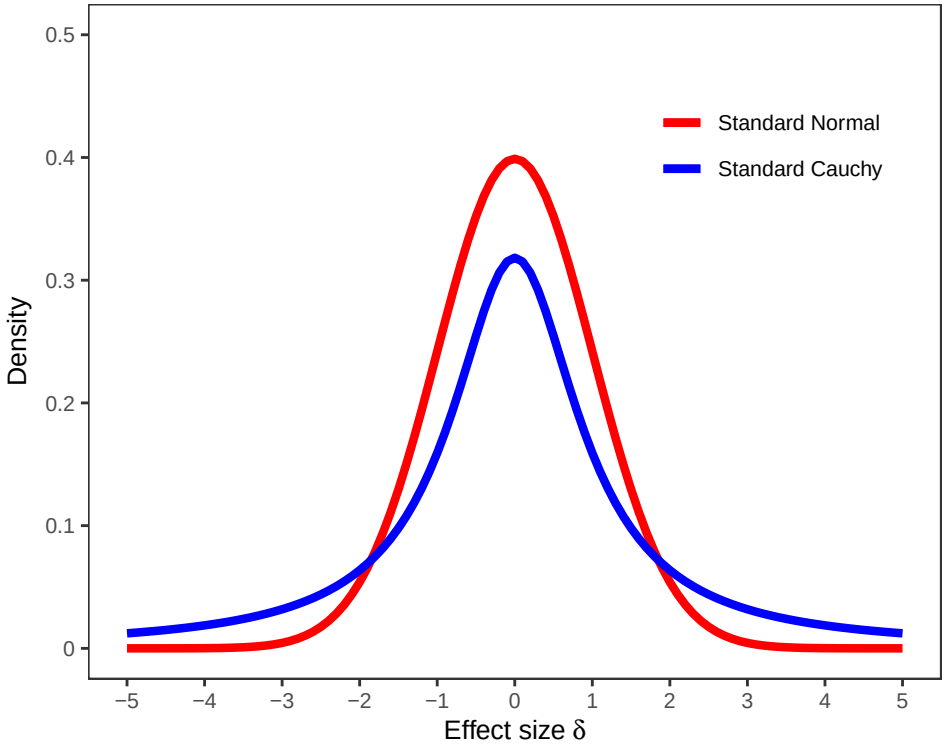

Figure 1. Comparison of the standard Normal probability density function and the standard Cauchy probability density function. The standard Normal and the standard Cauchy distributions are shown in red and blue, respectively.

1 has the advantage that the resulting Bayes factor is 1 in case of completely uninformative data. Moreover, the Bayes factor approaches infinity (or 0) for decisive data (Jeffreys, 1961). Still, by varying the Cauchy scale parameter, we can set a different emphasis on the credibility of a range of effect sizes. The larger the Cauchy prior scale, the stronger the support in favour of $\mathcal{H}_0$, given a fixed obtained effect size (e.g., Rouder et al., 2009). That is, with a larger scale, the Cauchy probability density function is more spread out and encompasses a wider range of values. More recently, a Cauchy prior scale of $r = 1/\sqrt{2}$ is used more and more often as a standard, as evident in the BayesFactor software (Morey & Rouder, 2018) and the point-and-click software JASP (JASP Team, 2018) and Jamovi (The jamovi project, 2019). Indeed, this value is for now also the standard in our baymedr software and is used in our simulations, as well. Nevertheless, objective priors are often critiqued on the grounds that they might not be that objective, after all, and that they might not follow the prior knowledge of the researcher (e.g., Kruschke & Liddell, 2018a; Tendeiro & Kiers, 2019). However, even proponents of objective priors acknowledge that more informed priors should be utilised if relevant knowledge is available (cf., e.g., Gronau et al., 2019; Rouder et al., 2009).

Implementation and Usage in baymedr

With our baymedr software (BAYesian inference for MEDical designs in R) written in R (R Core Team, 2019), one can easily calculate Bayes factors for superiority, non-inferiority, and equivalence designs. The user has the choice to supply raw data (i.e., arguments 'x' and 'y') or summary statistics (i.e., arguments 'n_x' and 'n_y' for sample sizes, 'mean_x' and 'mean_y' for means, and 'sd_x' and 'sd_y' for standard deviations [or, instead of 'sd_x' and 'sd_y', 'ci_margin' for the confidence interval margin of the difference in group means and 'ci_level' for the confidence level]). The latter option might by convenient when existing findings shall be reanalysed and one has to use the descriptive statistics that are reported in an article. Throughout, function arguments that have 'x' as a name or as a suffix refer to the control condition and those with 'y' as a name or suffix to the experimental condition.

In the following, we will restate the hypotheses for the superiority, non-inferiority,

and equivalence designs and explain how the corresponding Bayes factors can be calculated with baymedr. Again, we refer the reader to the article by van Ravenzwaaij et al. (2019) for the mathematical details (see also Gronau et al., 2019).

The Bayesian superiority test. The superiority test has a point null hypothesis $\mathcal{H}_0 : \delta = 0$ and a two-sided alternative hypothesis $\mathcal{H}_1 : \delta \neq 0$. Additionally, we can also define a one-sided alternative hypothesis: To do justice to the name of the superiority test, we can define $\mathcal{H}_1 : \delta > 0$ in the case where higher values on the measure of interest correspond to 'superiority' and $\mathcal{H}_1 : \delta < 0$ in the case where lower values represent 'superiority'.

Using baymedr, we can perform the Bayesian superiority test with the 'super_bf()' function. In order to comply with different ranges of plausible effect sizes under the alternative hypothesis, the Cauchy prior scale can be specified with the 'prior_scale' argument, although a default prior scale of $r = 1/\sqrt{2}$ is used if the 'prior_scale' argument is undefined. Depending on the research setting, low or high scores on the measure of interest represent 'superiority', which can be specified by the argument 'direction'. Moreover, since there are diverging practices on whether to conduct a one- or two-sided test for the superiority design, the user can make a decision for a one- or two-sided test by using the argument 'alternative'. After the test is run, a straightforward summary of all the specifications for the superiority test, a repetition of the corresponding hypotheses, and the resulting Bayes factor are printed in the console. Since we seek to find evidence for the alternative hypothesis (superiority), the Bayes factor quantifies evidence for $\mathcal{H}_1$ relative to $\mathcal{H}_0$ (i.e., BF_{10}).

The Bayesian non-inferiority test. The non-inferiority test is very similar to the superiority test. For the non-inferiority test we have $\mathcal{H}_0 : \delta < -nim$ and $\mathcal{H}_1 : \delta > -nim$ in the case where higher values on the measure of interest represent 'non-inferiority'; and we have $\mathcal{H}_0 : \delta > nim$ and $\mathcal{H}_1 : \delta < nim$ in the case where lower values represent 'non-inferiority'. Here, nim represents the non-inferiority margin. Note that, in contrast to the superiority test, we do not have a point null hypothesis for the non-inferiority test (cf. van Ravenzwaaij et al., 2019).

The Bayes factor can be calculated with the 'infer_bf()' function. As for the superiority test, the user can set the Cauchy prior scale with the 'prior_scale' argument, in order to set the desired emphasis on the magnitudes of effects under the alternative hypothesis. The value for the non-inferiority margin in unstandardised units can be specified with the 'ni_margin' argument. Lastly, depending on whether higher or lower values on the measure of interest represent 'non-inferiority', one of the options 'high' or 'low' should be determined for the argument 'direction'. Here again, we wish to determine the evidence in favour of $\mathcal{H}_1$; therefore, the evidence is expressed for $\mathcal{H}_1$ relative to $\mathcal{H}_0$ (i.e., BF_{10}). The output of the Bayesian non-inferiority test shows a user-friendly summary of the specifications and hypotheses, and displays the resulting Bayes factor.

The Bayesian equivalence test. In the equivalence test, we compare the null hypothesis $\mathcal{H}_0 : \delta < \Delta_L$ OR $\delta > \Delta_U$ with the alternative hypothesis $\mathcal{H}_1 : \delta > \Delta_L$ AND $\delta < \Delta_U$. In fact, these interval hypotheses are necessary for the frequentist equivalence test (see e.g., Meyners, 2012). However, the Bayesian equivalence test allows either an interval hypothesis or a point hypothesis for $\mathcal{H}_0$ (van Ravenzwaaij et al., 2019). Thus, with a point null hypothesis, the hypotheses become $\mathcal{H}_0 : \delta = 0$ and $\mathcal{H}_1 : \delta \neq 0$, which are the same as for the two-sided superiority test. Therefore, with a point null hypothesis, the equivalence and two-sided superiority tests are the same. The only difference is that the equivalence test standardly quantifies evidence in favour of $\mathcal{H}_0$ relative to $\mathcal{H}_1$. Still, the reciprocal of the resulting Bayes factor for the point equivalence test is equal to the Bayes factor for the two-sided superiority test.

baymedr can be utilised to conduct a Bayesian equivalence test with the 'equiv_bf()' function. As usual, the user can specify the width of the Cauchy prior distribution with the 'prior_scale' argument. The desired equivalence interval can be specified in standardised units with the 'interval' argument. Several options are possible: A symmetric equivalence interval around $\delta = 0$ is selected in case one value is provided (e.g., interval = 0.3) or a vector of length two, containing the same two values for the lower and upper equivalence interval boundaries, is inserted (e.g.,

interval = c (0.3, 0.3)). In contrast, an asymmetric equivalence interval can be specified with a vector of length two (e.g., interval = c (−0.5, 0.3)). Importantly, the implementation of a point null hypothesis is achieved by using either interval = 0 or interval = c (0, 0), which also serves as the default specification. Similar to the superiority and non-inferiority tests, the output of the equivalence test shows the model specifications, the hypotheses, and the resulting Bayes factor in a straightforward format.

Extracting Bayes factors. As already mentioned, when the functions for the superiority, non-inferiority, and equivalence designs are directly evaluated (i.e., without assigning them to a variable), a summary of the corresponding analysis and the resulting Bayes factor are printed in the console. However, when the results of these Bayesian tests are assigned to a variable, all relevant information for the corresponding analysis is stored in an S4 object. In simple terms, S4 is a system for object-oriented programming within R (R Core Team, 2019; see Wickham, 2019, for an overview). In certain situations it might be desirable to further manipulate the Bayes factor that is stored in a slot within the S4 object (e.g., taking the reciprocal of the Bayes factor). To this end, the corresponding Bayes factor needs to be extracted from the stored S4 object. This can be done with the 'get_bf()' function, which takes the S4 object as input.

An example. To illustrate how baymedr can be used, we will provide one example of an equivalence test (i.e., 'equiv_bf()') that is applied to fictitious data. The functions for the superiority and non-inferiority tests (i.e., 'super_bf()' and 'infer_bf()', respectively) can be used quite similarly, so we will not provide explicit examples for these tests. Let us pretend that we want to reanalyse the results from a previous study. Since we do not have access to the raw data, we have to use the summary statistics that are provided in the article. The control group has a sample size of $n = 220$, a mean of $M = 5.7$, and a standard deviation of $SD = 3.4$; the experimental group has a sample size of $n = 190$, a mean of $M = 6.1$, and a standard deviation of $SD = 3.9$. Moreover, the fictitious authors of the fictitious article used a symmetric equivalence interval with a margin of 0.1. For our Bayesian reanalysis, we choose to adhere to the default Cauchy

prior scale of $r = 1/\sqrt{2}$.

To conduct the equivalence test with the given summary statistics and parameter values for the Cauchy prior scale and the equivalence interval in baymedr, we can input:

```
equiv_bf(n_x = 220, n_y = 190,
         mean_x = 5.7, mean_y = 6.1,
         sd_x = 3.4, sd_y = 3.9,
         prior_scale = 1 / sqrt(2),
         interval = 0.1)
```

Since our Cauchy prior scale of choice represents the default value in baymedr, it would not have been necessary to provide this argument; however, for purposes of illustration, we mentioned it explicitly in the function call. Further, note that only a single value has to be specified for the equivalence interval in case a symmetric interval is desired. Using this function call, we receive the following output:

```
******************************
Equivalence analysis
--------------------
Data: summary data
H0 (equivalence):     mu_y - mu_x > c_low AND mu_y - mu_x < c_high
H1 (non-equivalence): mu_y - mu_x < c_low OR mu_y - mu_x > c_high
Equivalence interval: Lower = -0.1; Upper = 0.1
Prior scale: 0.707


    BF01 (equivalence) = 8.61157
******************************
```

The first part of the output provides a concise and user-friendly summary of the Bayesian equivalence analysis. The type of test and the type of data is mentioned. Moreover, the null and alternative hypotheses are stated. Subsequently, the lower and upper boundaries of the equivalence interval and the Cauchy prior scale are displayed.

The last part presents the resulting Bayes factor for the equivalence hypothesis (i.e., $\mathcal{H}_0$) relative to the non-equivalence hypothesis (i.e., $\mathcal{H}_1$). To avoid any potential confusion, the hypothesis towards which the Bayes factor quantifies evidence is stated within brackets (e.g., in this case 'equivalence').

Interim Summary and Prelude to the Simulations

Superiority, non-inferiority, and equivalence designs are important methods to analyse data for clinical trials. A majority of these tests is analysed with the frequentist approach (Chavalarias et al., 2016), in particular with NHSTs and corresponding p-values. Unfortunately, this implies several limitations: Evidence in favour of the null hypothesis cannot be quantified (e.g., Gallistel, 2009; van Ravenzwaaij et al., 2019; Wagenmakers, 2007; Wagenmakers et al., 2018) and researchers need to adhere very strictly to a predetermined sampling plan (Rouder, 2014; Schönbrodt & Wagenmakers, 2018; Schönbrodt et al., 2017). Moreover, in some situations, the frequentist non-inferiority and equivalence designs are accompanied with interpretational ambiguities.

Our baymedr software enables researchers to straightforwardly calculate Bayes factors for superiority, non-inferiority, and equivalence designs (see also van Ravenzwaaij et al., 2019). The Bayesian approach remedies all of the aforementioned shortcomings. We believe that these theoretical advantages of Bayesian inferences are so compelling, that the general adoption of these tools is warranted. Nevertheless, it is still to be examined how the frequentist and Bayesian approaches compare in actual performance. For this purpose, we conducted simulations for the equivalence and non-inferiority designs, which will be described next. Note that we did not conduct simulations for the superiority design, since similar simulations were already performed by van Ravenzwaaij and Ioannidis (2018).

Receiver Operating Characteristic Curves

Before we describe the simulations for the equivalence and the non-inferiority designs, we briefly explain what a receiver operating characteristic (ROC) analysis is and how it can be utilised (see, e.g., Fawcett, 2006; Macmillan & Creelman, 2005;

Stanislaw & Todorov, 1999, for thorough but accessible introductions; see also Zweig & Campbell, 1993, for medical applications). This is important because the ROC methodology is an essential part of the simulations.

Suppose we have instances or cases belonging to one of two classes, which we generically label as 'positive' and 'negative'; these are the true states. We have a binary classifier or diagnostic system, which predicts whether a given instance belongs to the positive or negative category. In case the output of the classifier is continuous, a threshold needs to be defined that marks the boundary between positive and negative predictions. For a given instance and classifier, four outcomes are possible: If the true state of an instance is positive and the classifier makes a positive prediction, this is termed as a true positive; if the prediction is negative, it is a false negative. In turn, if the true state of an instance is negative and the classifier makes a positive prediction, this is a false positive; if the prediction is negative, it is a true negative. With multiple instances, we can count the occurrences of these four possible outcomes and summarise them in a 2×2 matrix, where one dimension encompasses the two true states and the other dimension the two predicted states. Using this matrix, we can calculate the true positive rate, which is the proportion of correct predictions for truly positive instances:

$$TPR = \frac{\sum\limits_{i} TP_i}{\sum\limits_{i} TP_i + \sum\limits_{i} FN_i}. \tag{11}$$

Here, i is an index of the individual instances, TPR is the true positive rate, TP are true positives, and FN are false negatives. Similarly, we can calculate the false positive rate, which is the proportion of incorrect predictions for truly negative instances:

$$FPR = \frac{\sum\limits_{i} FP_i}{\sum\limits_{i} FP_i + \sum\limits_{i} TN_i}. \tag{12}$$

Here, i is an index of the individual instances, FPR represents the false positive rate, FP the false positives, and TN the true negatives. Importantly, these true and false positive rates only summarise the performance of the classifier for one specific decision

threshold. When we select other decision thresholds, different true and false positive rates will be obtained.

A ROC graph is a two-dimensional display that summarises the discriminatory ability of a binary classifier. The false positive rate is represented by the x-axis and the true positive rate by the y-axis. When we apply various decision thresholds to a classifier, the resulting true and false positive rates can be plotted in the ROC space, together forming a ROC curve. The antidiagonal line, from the bottom left to the top right, indicates chance performance because the true and false positive rates are equal. The closer the classifier is to the top left corner, the better its performance. ROC curves excellently illustrate the trade-off between the true and false positive rates of a classifier when the decision threshold is varied. The more conservative the decision threshold for making positive predictions, the lower the true and false positive rates. Similarly, a liberal decision threshold for positive predictions is associated with higher true and false positive rates. This trade-off is crucial in various applications. For instance, for medical tests, practitioners must decide whether it is worse to diagnose healthy people as ill (liberal; high false and true positive rates) or to diagnose ill people as healthy (conservative; low false and true positive rates).

Simulations for the Equivalence Design

Method

Data generation. We generated multiple data sets that imitated two-condition (e.g., control and experimental) independent-samples experiments. These data sets differed along two dimensions that were fully crossed. First, we varied the true population effect size between the experimental and control conditions. We selected effect sizes of $\delta = 0.25$, $\delta = 0.5$, and $\delta = 0.75$ (non-zero-effect; true non-equivalence), roughly in accordance with Cohen's (1988) approximate guidelines for judging the magnitude of an effect (small, medium, and large, respectively). Second, the total sample size was either $n = 50$, $n = 100$, $n = 500$, $n = 1,000$, or $n = 5,000$, with an equal number of cases in both conditions. The individual combinations of the true population effect size and sample size parameters yielded 15 different types of data sets.

We generated 1,000 repetitions of each data set type, forming 15 data set assemblies. Importantly, however, each data set assembly additionally contained 1,000 data sets with a true population effect size of $\delta = 0$ (zero-effect; true equivalence) and the same sample size as the other data sets within the corresponding data set assembly, resulting in 2,000 data sets within each data set assembly. In total, therefore, we generated 30,000 data sets.

The values for each condition within each data set of a data set assembly were drawn from a Normal distribution, differing only in the mean argument:

$$
\begin{aligned}
c &\sim N\left(0,\ 1\right) \\
e_0 &\sim N\left(0,\ 1\right) \\
e_\delta &\sim N\left(\delta,\ 1\right),
\end{aligned}
$$

where c refers to the control condition, e_0 to the experimental condition with a zero-effect, and e_δ to the experimental condition with a non-zero-effect. The mathematical notation $\sim N\left(a,\ b\right)$ indicates that data values were drawn from a Normal distribution defined by a mean of a and a standard deviation of b.

In summary, we generated data sets varying in the true population effect size and the sample size, yielding 15 different types of data sets. Each type of data set was repeated 1,000 times, forming 15 data set assemblies. Further, each data set assembly also contained 1,000 data sets with zero-effects, resulting in a total of 2,000 data sets within each data set assembly. Data values for each condition within each data set of each data set assembly were drawn from Normal distributions with varying means.

Calculation of p-values and Bayes factors. Frequentist equivalence tests (i.e., TOST; Lakens, 2017; Lakens et al., 2018; Schuirmann, 1987) were conducted using the TOSTER software (Lakens, 2017) and Bayesian equivalence tests were conducted using our baymedr software. Both packages are written in R (R Core Team, 2019).

Each data set was analysed both by the frequentist and the Bayesian equivalence tests. The frequentist approach consists of conducting two one-sided t-tests. As soon as

one of these two t-tests results in a p-value that is higher than the predefined type I error rate (α), the null hypothesis of non-equivalence cannot be rejected (Meyners, 2012). Therefore, only the largest of the two p-values is of importance and was considered for further analyses. Here, we assumed unequal variances between the control and the experimental conditions. The Bayesian equivalence tests were conducted with a Cauchy prior and a corresponding scale parameter of $r = 1/\sqrt{2}$ for the effect size under the alternative hypothesis (Jeffreys, 1961; Liang et al., 2008; Rouder et al., 2009). Crucially, the resulting Bayes factors quantified evidence in the direction of the null hypothesis of equivalence (i.e., BF_{01}).

We selected four symmetrical equivalence intervals with margins of $0.05SD$, $0.1SD$, $0.15SD$, and $0.2SD$, resulting in four p-values and four Bayes factors for each data set. Hence, within each data set assembly, we calculated 8,000 p-values and 8,000 Bayes factors, 2,000 of each for each specification of the equivalence interval. Unfortunately, we could not include a point null hypothesis (i.e., an equivalence margin of $0SD$) in our analyses because, in contrast to the Bayesian approach, the frequentist equivalence test is not able to implement point null hypotheses (Meyners, 2012; van Ravenzwaaij et al., 2019). The resulting p-values and Bayes factors formed the basis for subsequent ROC analyses.

ROC analysis. We conducted a ROC analysis for each combination of true population effect size, sample size, and equivalence interval, yielding 60 ROC analyses. Thus, for each ROC analysis, 2,000 p-values and 2,000 Bayes factors were of relevance. Each of the 2,000 p-values was compared against three thresholds: $\alpha = .05$, $\alpha = .01$, and $\alpha = .005$. If the p-value was lower than α (i.e., $p < \alpha$), the test predicted a positive result, corresponding to equivalence. If the p-value was higher than α (i.e., $p > \alpha$), the results of the test were ambiguous, neither predicting equivalence nor non-equivalence, because the frequentist approach is not able to quantify evidence for the null hypothesis (Gallistel, 2009; van Ravenzwaaij et al., 2019; Wagenmakers, 2007; Wagenmakers et al., 2018). Similarly, each of the 2,000 Bayes factors was compared against three thresholds: $BF_{thr} = 3$, $BF_{thr} = 10$, $BF_{thr} = 30$. These Bayes factor thresholds correspond to the

approximate thresholds for anecdotal ($1 < BF < 3$), moderate ($3 < BF < 10$), and strong ($10 < BF < 30$) evidence (Jeffreys, 1961; Lee & Wagenmakers, 2013). As already mentioned, the Bayes factors obtained from the simulations quantify evidence in the direction of the null hypothesis (equivalence). Hence, if the Bayes factor was higher than BF_{thr}, the test predicted a positive result, corresponding to equivalence. However, if the Bayes factor was lower than BF_{thr}, the test result was ambiguous, neither predicting equivalence nor non-equivalence. Note that, in principle, we can quantify evidence both for the null (equivalence) and the alternative (non-equivalence) hypotheses with the Bayes factor. However, since we only defined a threshold for equivalence (BF_{thr}) but not for non-equivalence ($1/BF_{thr}$), any Bayes factor that is lower than the threshold for equivalence is treated as providing ambiguous evidence.

The combination of the true state (equivalence [zero-effect] vs. non-equivalence [non-zero-effect]) and the predictions/classifications (equivalence vs. ambiguous) yielded two 2×2 matrices with four possible outcomes, one matrix for the frequentist and one for the Bayesian approach. If the true state was equivalence and the prediction was equivalence, it was counted as a true positive; if the prediction was ambiguous, it was counted as a false negative. Similarly, if the true state was non-equivalence and the prediction was ambiguous, it was counted as a true negative; if the prediction was equivalence, it was counted as a false positive. Based on these matrices, we calculated the true positive rate (equation 11) and the false positive rate (equation 12).

Results and Discussion

The results of the ROC simulations for the equivalence test for the effect sizes of $\delta = 0.25$, $\delta = 0.5$, and $\delta = 0.75$ are shown in Fig 2, 3, and 4, respectively. The false positive rate is plotted on the x-axis and the true positive rate on the y-axis. Plot panels are defined by the total sample size (rows) and the equivalence interval (columns). Results of the Bayesian and frequentist tests are plotted in red and blue, respectively. The three shapes represent different decision thresholds, with an increasing degree of conservativeness as we switch from squares to circles to triangles. Data that fall on the antidiagonal line, starting at the bottom left and ending at the top right (not

shown), indicate chance performance. The closer the data approach the top left corner, where the true positive rate is at a maximum and the false positive rate at a minimum, the better the performance of the classifier.

The results reveal a familiar pattern or mathematical rule in ROC analyses: There

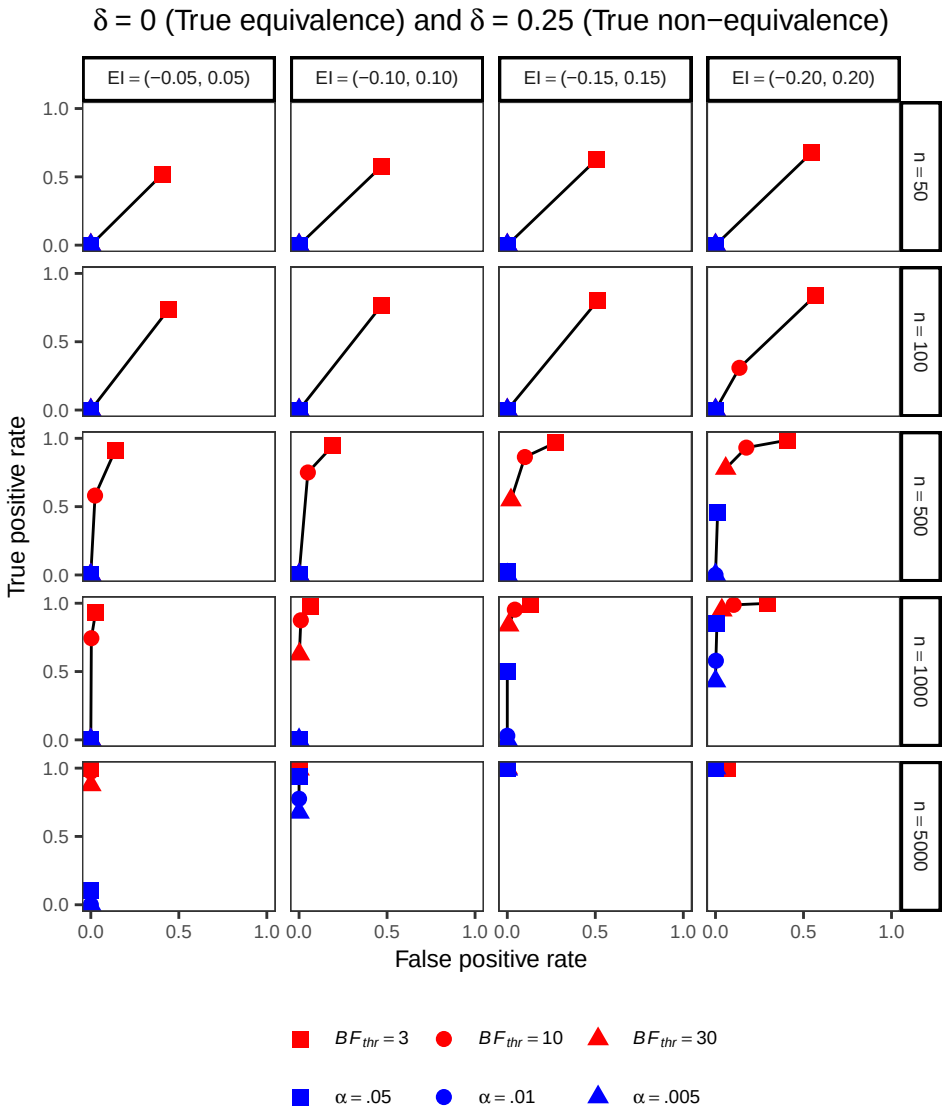

Figure 2. ROC curves for the equivalence test with true population effect sizes of $\delta = 0$ in case of true equivalence and $\delta = 0.25$ in case of true non-equivalence between the experimental and control conditions. A true positive corresponds to classifying truly equivalent conditions as equivalent; a false positive corresponds to classifying truly non-equivalent conditions as equivalent. The rows represent different sample sizes (n) and the columns represent different equivalence intervals (EI). Bayesian results are printed in red and frequentist results in blue. Shapes represent different decision thresholds with increasing conservativeness, moving from squares to circles to triangles.

is a trade-off between the true and false positive rates when shifting the decision criterion. The more conservative the decision threshold, the lower the false positive rate but also the true positive rate.

The frequentist equivalence test performed very similar across the three true

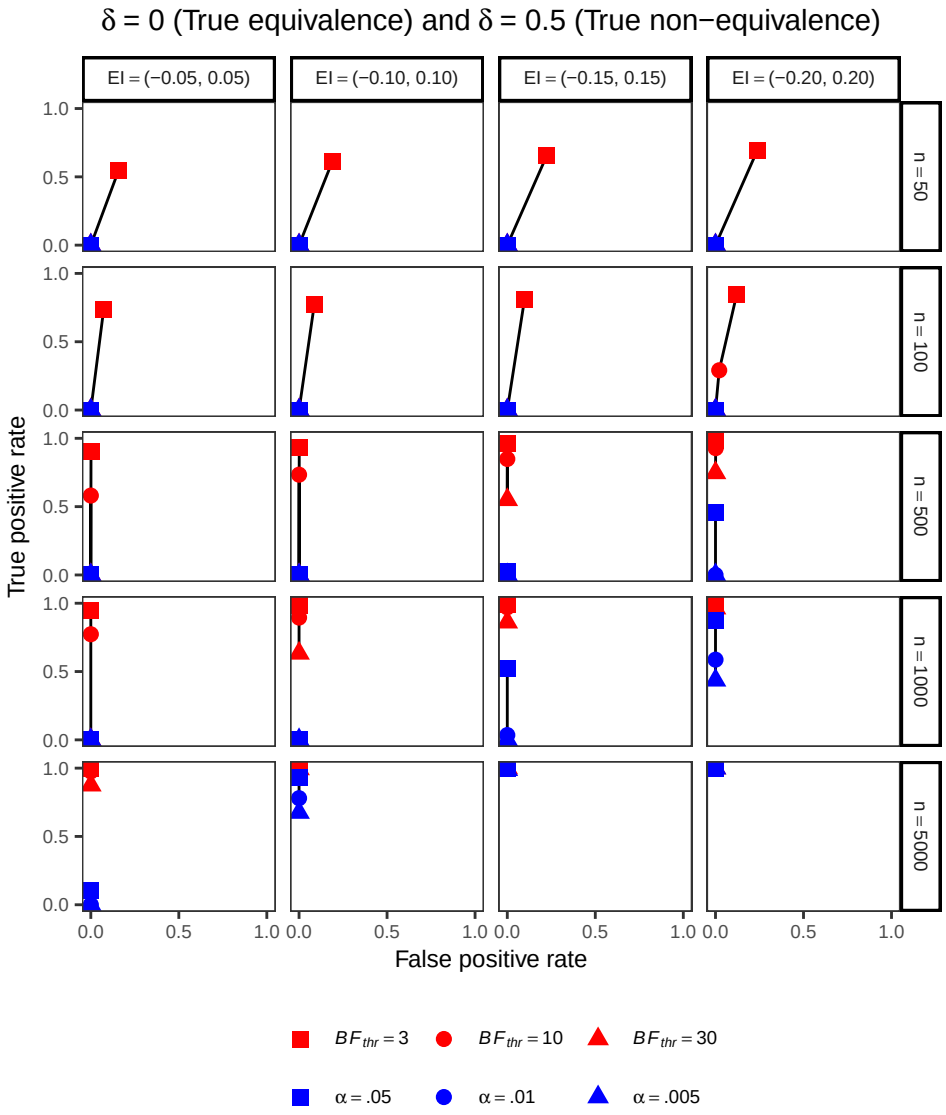

Figure 3. ROC curves for the equivalence test with true population effect sizes of $\delta = 0$ in case of true equivalence and $\delta = 0.5$ in case of true non-equivalence between the experimental and control conditions. A true positive corresponds to classifying truly equivalent conditions as equivalent; a false positive corresponds to classifying truly non-equivalent conditions as equivalent. The rows represent different sample sizes (n) and the columns represent different equivalence intervals (EI). Bayesian results are printed in red and frequentist results in blue. Shapes represent different decision thresholds with increasing conservativeness, moving from squares to circles to triangles.

population effect sizes for non-equivalent data sets. Across all true population effect sizes, sample sizes, equivalence intervals, and decision thresholds, the frequentist approach was characterised by a false positive rate of zero or almost zero. At the same time, however, its ability to classify conditions that were truly equivalent was extremely

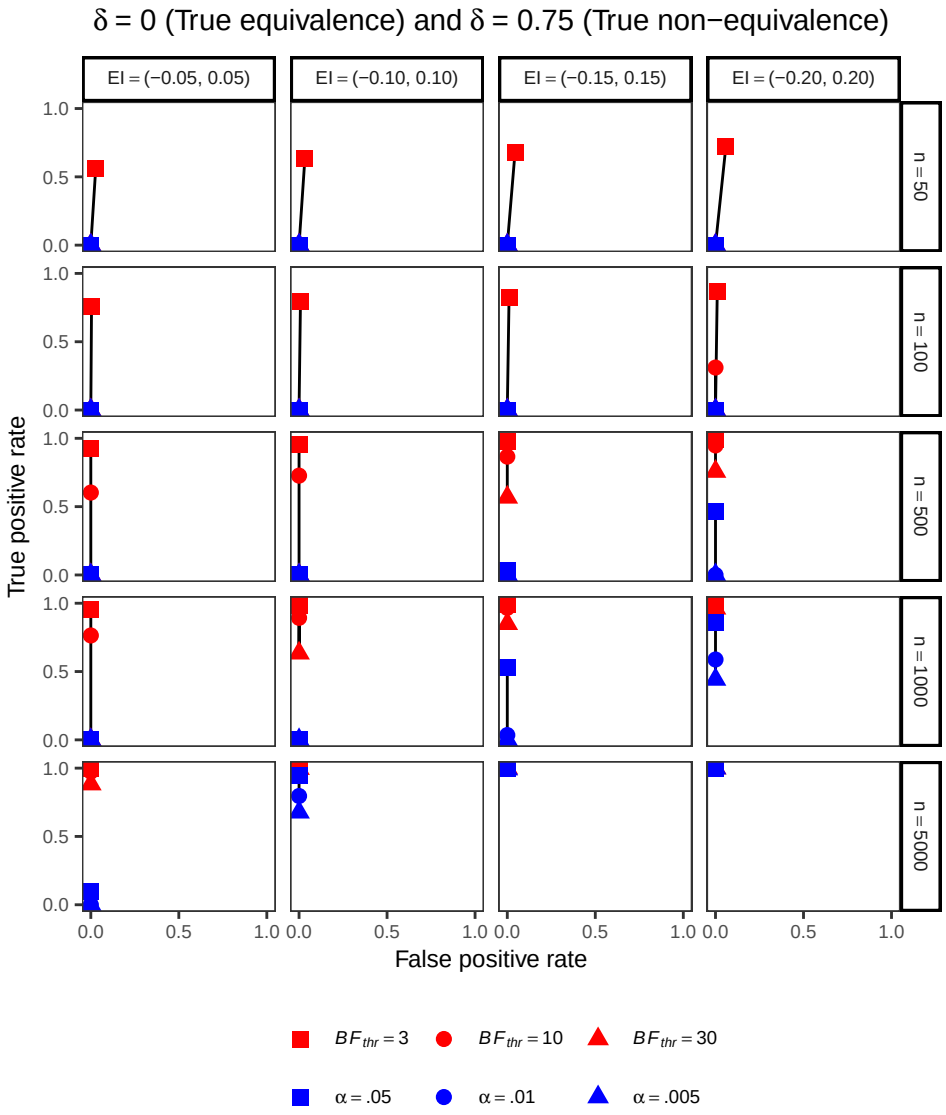

Figure 4. ROC curves for the equivalence test with true population effect sizes of $\delta = 0$ in case of true equivalence and $\delta = 0.75$ in case of true non-equivalence between the experimental and control conditions. A true positive corresponds to classifying truly equivalent conditions as equivalent; a false positive corresponds to classifying truly non-equivalent conditions as equivalent. The rows represent different sample sizes (n) and the columns represent different equivalence intervals (EI). Bayesian results are printed in red and frequentist results in blue. Shapes represent different decision thresholds with increasing conservativeness, moving from squares to circles to triangles.

low in cases of relatively small sample sizes and relatively narrow equivalence intervals, as evident by true positive rates very close to zero. This pattern of true and false positive rates of almost zero for small sample sizes and narrow equivalence intervals indicates that the prediction performance of the frequentist equivalence test was at chance level. Nevertheless, for higher sample sizes and broader equivalence intervals, the true positive rate markedly improved. With the highest sample size of $n = 5,000$ and an equivalence margin of either $0.15SD$ or $0.2SD$, the frequentist test reached nearly perfect classification performance.

The performance of the Bayesian equivalence test was largely influenced by the true population effect size in non-equivalent data sets. The larger the true population effect size, the better the performance. This was especially apparent for smaller sample sizes. In contrast to the frequentist equivalence test, the false positive rate was more variable and generally higher with the Bayesian approach, which was especially apparent with smaller true population effect sizes. However, this is compensated by a notable true positive rate in all panels for all true population effect sizes. Nevertheless, the performance is near chance in case of a true population effect size of $\delta = 0.25$ and a sample size of $n = 50$. For the other effect sizes and all other panels, there was at least an acceptable classification performance, which drastically increased with larger sample sizes and broader equivalence intervals.

Comparing the two approaches, the general pattern is that the Bayesian approach performed notably better than the frequentist approach, as evident by smaller distances to the upper left corner. Impressively, this is apparent in practically all panels across all true population effect sizes. The Bayesian equivalence test required far less cases to reach a reasonable classification performance than the frequentist approach, which is particularly important for expensive and potentially harmful clinical trials. In other words, the frequentist equivalence test had an extremely low power to detect truly equivalent conditions in case of small and mediocre sample sizes and narrow equivalence intervals. Of course, a remedy for this shortcoming would be to collect data for even more cases. Oftentimes, however, this is not viable, since sampling an increasing

number of cases is very expensive and time-consuming.

Exploration of Proper Bayesian Equivalence Tests

It is important to stress that in reality the Bayesian procedure for equivalence tests is not performed the same way as in the simulations of this article. Fig 5 illustrates this dichotomy by means of an example of the Bayesian equivalence test with a decision threshold of $BF_{thr} = 10$, represented by the changing red and blue line. The way the Bayesian equivalence test was conducted in the simulations was that we concluded equivalence if the Bayes factor was higher than 10; if the Bayes factor was lower than 10 we obtained ambiguous evidence, favouring neither equivalence nor non-equivalence. This procedure is illustrated by the red arrows and annotations in Fig 5. Mathematically, however, any Bayes factor larger than 1 presents evidence in favour

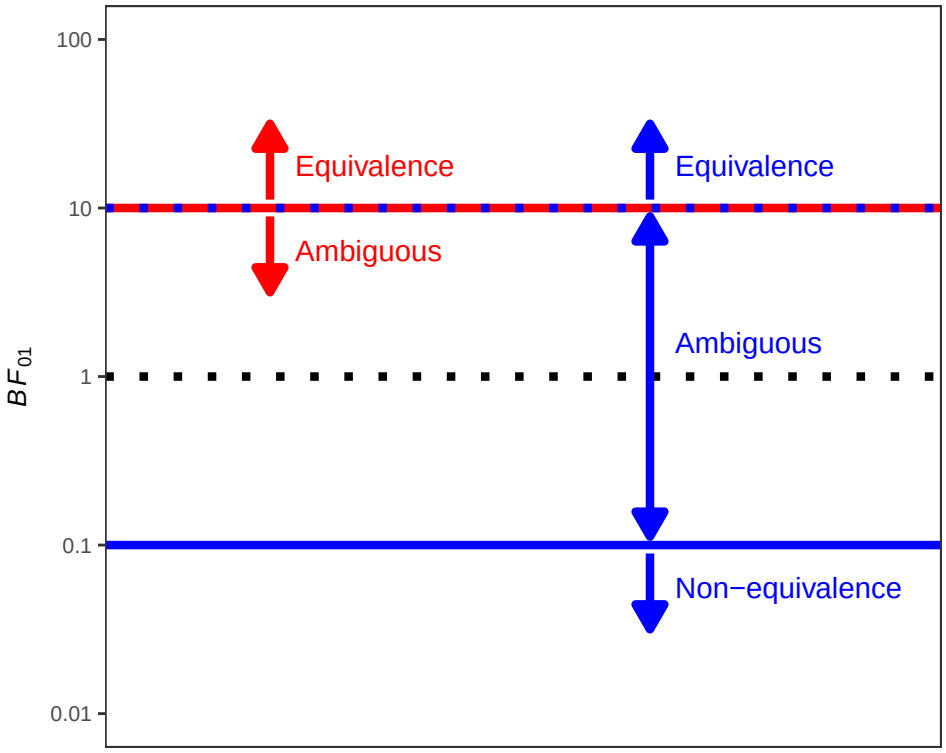

Figure 5. Contrast improper Bayesian decision making in our simulations with proper realistic Bayesian decision making, using the equivalence design with a threshold of $BF_{thr} = 10$ as an example. In our simulations, we would have concluded equivalence if the Bayes factor was larger than 10 (red arrows). In contrast, proper Bayesian decision making would conclude equivalence if the Bayes factor is larger than 10 and non-equivalence if the Bayes factor is smaller than 0.1; any Bayes factor between 0.1 and 10 would correspond to ambiguous or inconclusive evidence (blue arrows).

of equivalence and any Bayes factor smaller than 1 provides evidence for non-equivalence, no matter how little the evidence is. To obtain a sufficient amount of evidence, the Bayesian data analyst could define thresholds, which the Bayes factor has to reach in order to conclude equivalence or non-equivalence. In the present example, the threshold for equivalence would be 10 and the threshold for non-equivalence 0.1 (i.e., the reciprocal of 10). Consequently, in a real setting, the Bayesian data analyst would conclude equivalence if the Bayes factor is higher than 10 and non-equivalence if

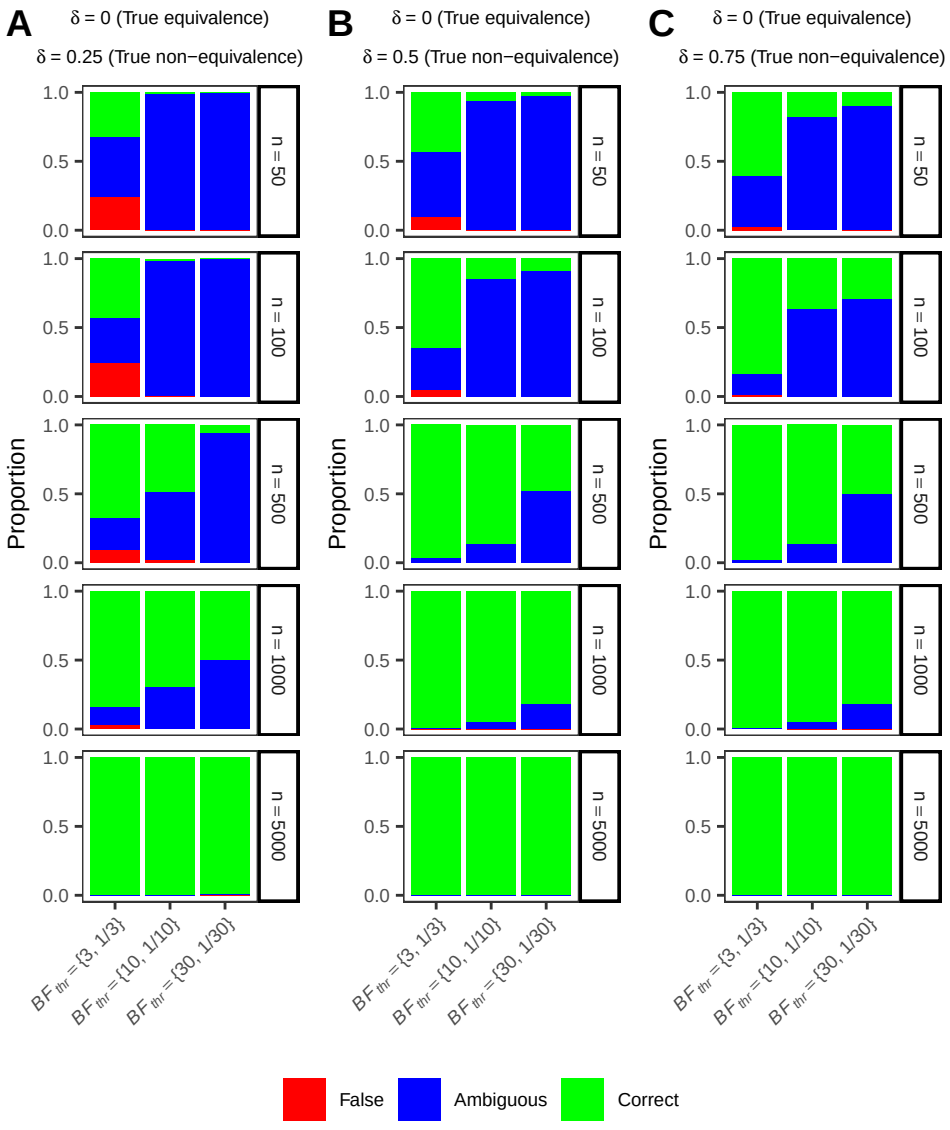

Figure 6. Exploration of correct, false, and ambiguous decisions for the equivalence design, using proper Bayesian decision making instead of binary decision making, as used in our simulations. The exemplary results correspond to an equivalence margin of $0.1SD$.

the Bayes factor is lower than 0.1 (in contrast to our simulations, where a Bayes factor that is smaller than 0.1 is treated as providing ambiguous evidence). Crucially, however, the analyst would judge any Bayes factor between 0.1 and 10 as ambiguous or inconclusive, requiring more evidence for a solid decision (see the blue arrows and annotations in Fig 5).

We explored the proportions of correct, false, and ambiguous decisions that would have been made in our simulations for the Bayesian equivalence design if we used proper Bayesian decision making, as outlined above. To limit the amount of visualisations, we only examined the results for an equivalence margin of $0.1SD$ as exemplification, which we chose because it was of mediocre size in our original set of equivalence margins (i.e., $0.05SD$, $0.1SD$, $0.15SD$, and $0.2SD$). A correct decision was made when the Bayes factor was larger than a threshold (i.e., $BF_{thr} = 3$, $BF_{thr} = 10$, or $BF_{thr} = 30$) and the conditions were truly equivalent *or* when the Bayes factor was smaller than the reciprocal of a threshold (i.e., $BF_{thr} = 1/3$, $BF_{thr} = 1/10$, or $BF_{thr} = 1/30$) and the conditions were truly non-equivalent. Conversely, a false decision was made when the Bayes factor was larger than a threshold and the conditions were truly non-equivalent *or* when the Bayes factor was smaller than the reciprocal of a threshold and the conditions were truly equivalent. Ambiguous evidence was obtained when the resulting Bayes factor lay between the threshold and the reciprocal of the threshold.

The panels A, B, and C in Fig 6 show the results for true population effect sizes of $\delta = 0.25$, $\delta = 0.5$, and $\delta = 0.75$ in case of truly non-equivalent data sets, respectively. The proper Bayesian equivalence test performed very well. There were non-negligible proportions of false predictions but they mostly occurred in the cases where very lenient evidence thresholds (e.g., $BF_{thr} = 3$) and small sample sizes were chosen and the true population effect size was close to the equivalence interval boundary. With small sample sizes and conservative decision criteria, the majority of findings were ambiguous.

Simulations the Non-Inferiority Design

Method

The general structure of the simulations for the non-inferiority design was very similar to the simulations for the equivalence design. Therefore, commonalities are only mentioned very briefly; in turn, differences are described in detail. We refer readers to the Method section of the equivalence simulations for a full description.

Data generation. As in the simulations for the equivalence design, the generated data sets imitated two-condition independent-samples experiments. While the set of sample sizes was the same as before (i.e., $n = 50$, $n = 100$, $n = 500$, $n = 1,000$, and $n = 5,000$), we changed the true population effect sizes for truly inferior data sets to $\delta = -0.3$, $\delta = -0.4$, and $\delta = -0.5$, resulting in 15 different types of data sets. Each type of data set was repeated 1,000 times to form a data set assembly. In addition, 1,000 data sets with a true population effect size of $\delta = 0$ (zero-effect; true non-inferiority) were added to each data set assembly. Values for the control condition and the experimental condition with a zero-effect were drawn from a Normal distribution with a mean of $\mu = 0$ and a standard deviation of $\sigma = 1$, whereas the values for the experimental condition with a non-zero-effect were sampled from a Normal distribution with a mean defined by the true population effect size and a standard deviation of $\sigma = 1$. In summary, the data generation process differed only in the specification of the true population effect sizes.

Calculation of p-values and Bayes factors. Frequentist non-inferiority tests were conducted by performing two-condition independent-samples t-tests, using R (R Core Team, 2019). Bayesian non-inferiority tests were conducted using the baymedr software, written in R (R Core Team, 2019).

Each data set was analysed by both the frequentist and the Bayesian non-inferiority tests. The frequentist approach consisted of conducting one-sided t-tests, assuming unequal variances between the two conditions. The non-inferiority margin was incorporated by comparing the resulting effect size between the two conditions to the respective negative non-inferiority margin (instead of the conventional centre of the null

hypothesis at 0). The Bayesian non-inferiority tests were conducted with a Cauchy prior and a scale parameter of $r = 1/\sqrt{2}$ for the effect size under the alternative hypothesis (see e.g., Jeffreys, 1961; Liang et al., 2008; Rouder et al., 2009). A one-sided Bayes factor was calculated. The resulting Bayes factors quantified evidence in the direction of the alternative hypothesis of non-inferiority (i.e., BF_{10}).

We selected non-inferiority margins of $0.05SD$, $0.1SD$, $0.15SD$, and $0.2SD$, resulting in four p-values and four Bayes factors for each data set. Hence, within each data set assembly, we calculated 8,000 p-values and 8,000 Bayes factors, 2,000 of each for each specification of the non-inferiority margin. The resulting p-values and Bayes factors were used for the ROC analysis.

ROC analysis. We conducted a ROC analysis for each combination of true population effect size, sample size, and non-inferiority margin, yielding 60 ROC analyses. Thus, for each ROC analysis, 2,000 p-values and 2,000 Bayes factors were of relevance. Each of the 2,000 p-values was compared against three thresholds: $\alpha = .05$, $\alpha = .01$, and $\alpha = .005$. If $p < \alpha$, the test predicted a positive result (i.e., non-inferiority). If $p > \alpha$, the test result was ambiguous, neither predicting inferiority nor non-inferiority. Similarly, each of the 2,000 Bayes factors was compared against three thresholds: $BF_{thr} = 3$, $BF_{thr} = 10$, $BF_{thr} = 30$. If the Bayes factor was higher than BF_{thr}, the test predicted a positive result (i.e., non-inferiority). However, if the Bayes factor was lower than BF_{thr}, the test result was ambiguous. Similar to the simulations for the equivalence design, we only defined a threshold for non-inferiority (BF_{thr}) but not for inferiority ($1/BF_{thr}$). Therefore, any Bayes factor that is lower than the threshold for non-inferiority provides ambiguous evidence.

The combination of the true state (non-inferiority [zero-effect] vs. inferiority [non-zero-effect]) and the predictions/classifications (non-inferiority vs. ambiguous) yielded two 2×2 matrices, one for the frequentist and one for the Bayesian approach. If the true state was non-inferiority and the prediction was non-inferiority, it was counted as a true positive; if the prediction was ambiguous, it was counted as a false negative. Similarly, if the true state was inferiority and the prediction was ambiguous, it was

counted as a true negative; if the prediction was non-inferiority, it was counted as a false positive. Based on these matrices, we calculated the true positive and the false positive rates (see equations 11 and 12, respectively).

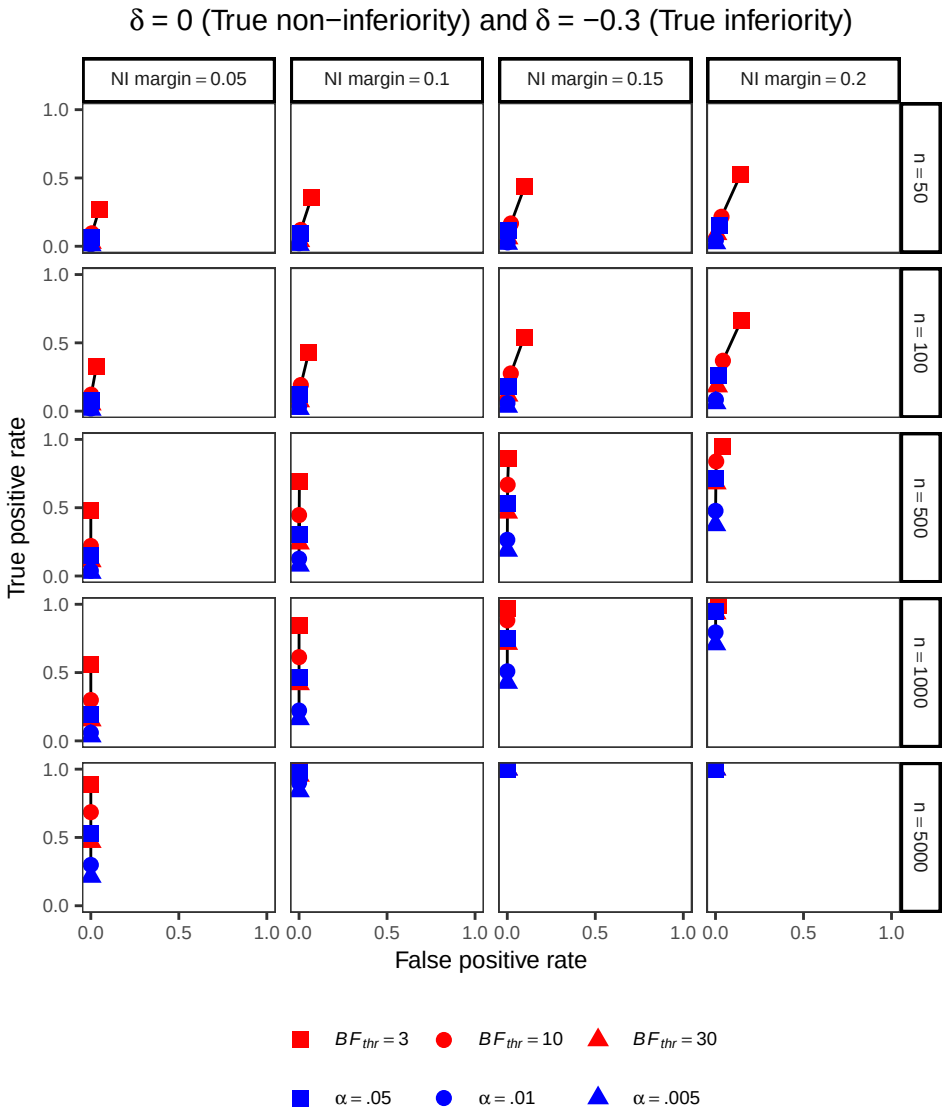

Figure 7. ROC curves for the non-inferiority test with true population effect sizes of $\delta = 0$ in case of true non-inferiority and $\delta = -0.3$ in case of true inferiority of the experimental condition compared to the control condition. A true positive corresponds to classifying a truly non-inferior experimental condition as non-inferior; a false positive corresponds to classifying a truly inferior experimental condition as non-inferior. The rows represent different sample sizes (n) and the columns represent different non-inferiority margins ($NI\ margin$). Bayesian results are printed in red and frequentist results in blue. Shapes represent different decision thresholds with increasing conservativeness, moving from squares to circles to triangles.

Results and Discussion

The results of the ROC simulations for the non-inferiority test for the effect sizes of $\delta = -0.3$, $\delta = -0.4$, and $\delta = -0.5$ are shown in Fig 7, 8, and 9, respectively. The plot layout is the same as in the plots for the equivalence design, except that columns now

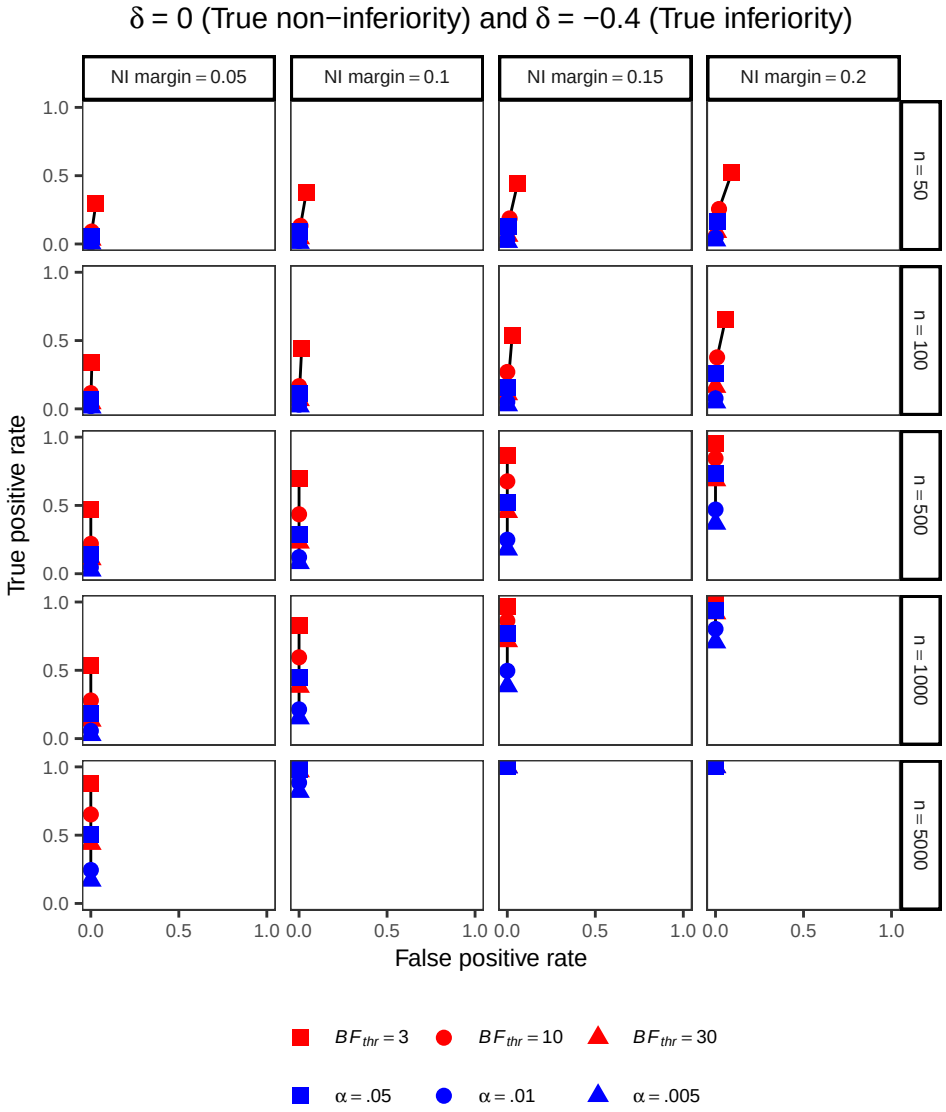

Figure 8. ROC curves for the non-inferiority test with true population effect sizes of $\delta = 0$ in case of true non-inferiority and $\delta = -0.4$ in case of true inferiority of the experimental condition compared to the control condition. A true positive corresponds to classifying a truly non-inferior experimental condition as non-inferior; a false positive corresponds to classifying a truly inferior experimental condition as non-inferior. The rows represent different sample sizes (n) and the columns represent different non-inferiority margins ($NI\ margin$). Bayesian results are printed in red and frequentist results in blue. Shapes represent different decision thresholds with increasing conservativeness, moving from squares to circles to triangles.

represent non-inferiority margins instead of equivalence intervals.

Here again, the results show the trade-off between the true and false positive rates: As the decision threshold gets more conservative, the false and true positive rates decrease, and vice versa.

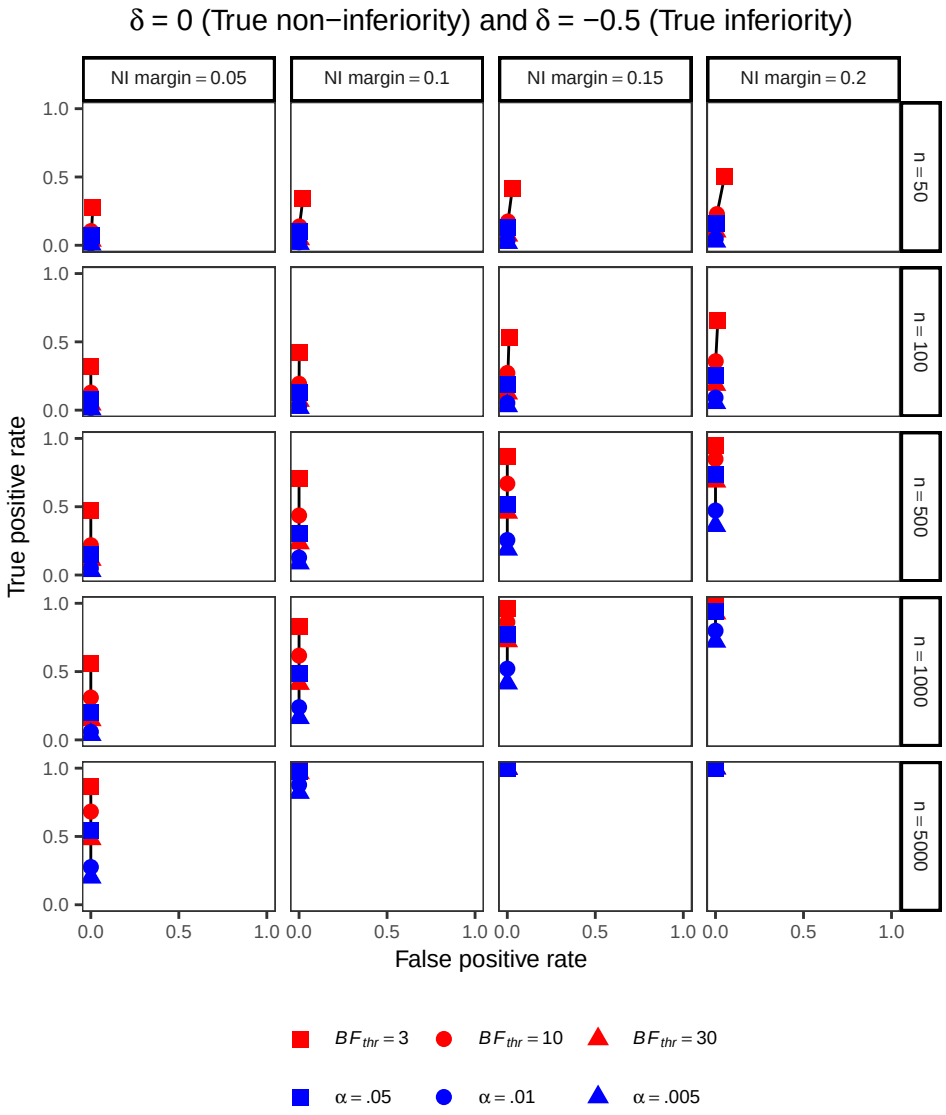

Figure 9. ROC curves for the non-inferiority test with true population effect sizes of $\delta = 0$ in case of true non-inferiority and $\delta = -0.5$ in case of true inferiority of the experimental condition compared to the control condition. A true positive corresponds to classifying a truly non-inferior experimental condition as non-inferior; a false positive corresponds to classifying a truly inferior experimental condition as non-inferior. The rows represent different sample sizes (n) and the columns represent different non-inferiority margins ($NI\ margin$). Bayesian results are printed in red and frequentist results in blue. Shapes represent different decision thresholds with increasing conservativeness, moving from squares to circles to triangles.

As was the case for the frequentist equivalence test, the frequentist non-inferiority test performed similar across our three specifications of the true population effect size for inferiority data sets. Again, the false positive rate was zero or almost zero across all true population effect sizes, sample sizes, non-inferiority margins, and decision thresholds. With small sample sizes, the true positive rate was very low, demonstrating that non-inferiority data sets were mostly incorrectly classified. However, in contrast to the frequentist equivalence test, the true positive rate readily and gradually increased as higher sample sizes and larger non-inferiority margins were used. Near-perfect classifications were reached with a sample size of $n = 1,000$ and the broadest non-inferiority margin ($0.2SD$) or a sample size of $n = 5,000$ and all non-inferiority margins, except for the narrowest ($0.05SD$).

Although the results of the Bayesian non-inferiority test varied across the three true population effect sizes for the inferiority data sets, this variability was less pronounced than for the Bayesian equivalence test. The larger the magnitude of the true population effect size, the better the performance. This was especially apparent for smaller sample sizes. For smaller sample sizes (i.e., $n = 50$ and $n = 100$), the false positive rate was higher compared to the frequentist approach, especially with a true population effect size of $\delta = -0.3$. This was compensated by higher true positive rates. The overall classification performance gradually improved with larger sample sizes and non-inferiority margins.

A comparison of the two approaches yields the general pattern that the Bayesian approach performed better than the frequentist approach. Impressively, this is apparent in practically all panels across all true population effect sizes. It is important to note, however, that this difference is less pronounced than in the equivalence simulations. Still it is apparent that the Bayesian non-inferiority test reached a rather good classification performance even with moderate sample sizes.

Exploration of Proper Bayesian Non-Inferiority Tests

As was the case in the simulations for the equivalence design, the decision making procedure for the Bayesian non-inferiority test in our simulations diverges from proper

Bayesian decision making (see Fig 5 for an illustration of this divergence for the equivalence design, which can likewise be applied to the non-inferiority design). Instead of interpreting Bayes factors higher than a threshold as providing evidence for non-inferiority and any Bayes factor below that threshold as providing ambiguous evidence, proper Bayesian decision making would conclude non-inferiority if the Bayes factor exceeds the non-inferiority threshold and inferiority if the Bayes factor is lower than the inferiority threshold (the reciprocal of the non-inferiority threshold); any Bayes

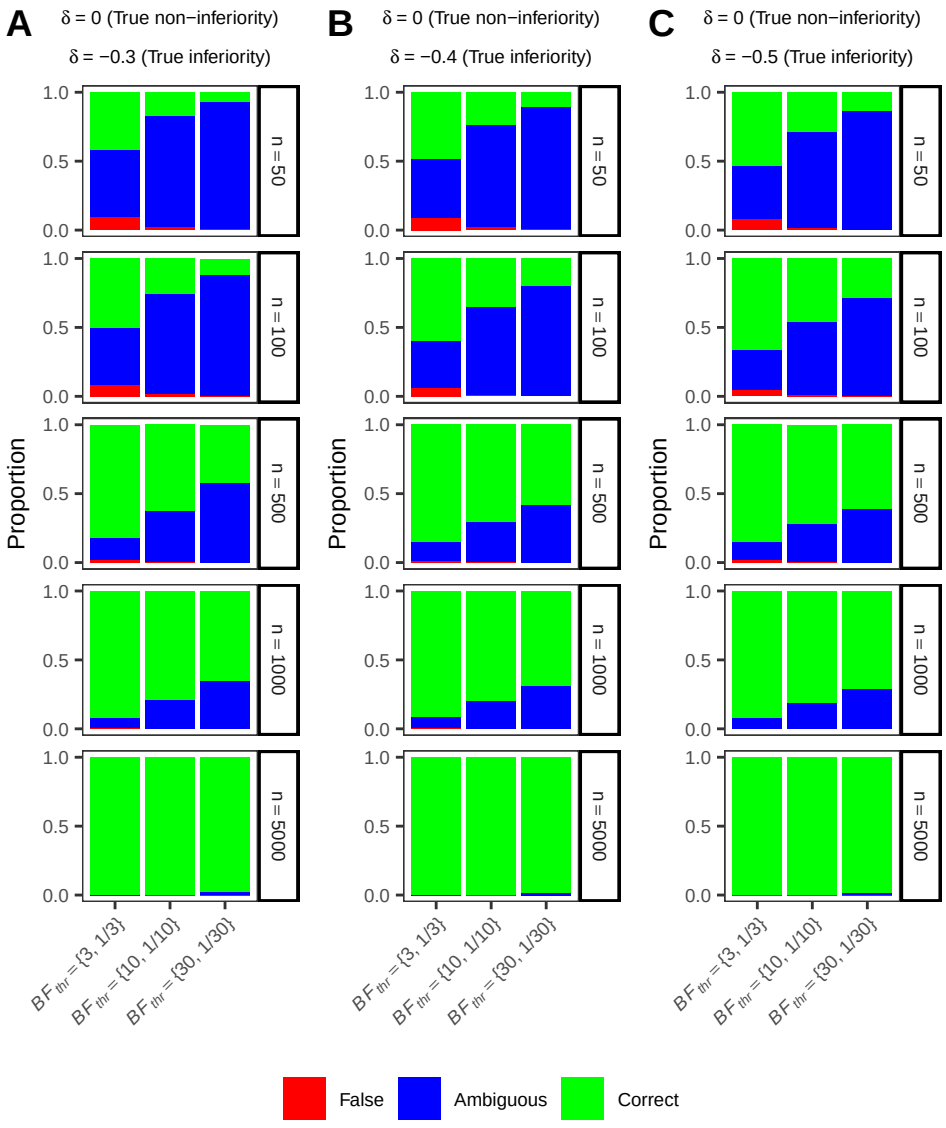

Figure 10. Exploration of correct, false, and ambiguous decisions for the non-inferiority design, using proper Bayesian decision making instead of binary decision making, as used in our simulations. The exemplary results correspond to a non-inferiority margin of $0.1SD$.

factor between the two threshold would provide ambiguous evidence.

We explored the proportion of correct, false, and ambiguous decisions that would have been made in our simulations for the Bayesian non-inferiority design if we used proper Bayesian decision making, as outlined above. Correct, false, and ambiguous decisions were defined as in the exploratory equivalence simulations. We used a non-inferiority margin of $0.1SD$ as an exemplification.

The panels A, B, and C in Fig 10 show the results for true population effect sizes of $\delta = -0.3$, $\delta = -0.4$, and $\delta = -0.5$ in case of truly inferior data sets, respectively. Although considerable proportions of false predictions were obtained for lenient evidence thresholds, small sample sizes, and when the true population effect size was close to the non-inferiority margin, the proper Bayesian non-inferiority was very accurate, in general. With small sample sizes and conservative decision criteria, the majority of findings were ambiguous.

General Discussion

In this article, we contrasted the common frequentist approach and the Bayesian approach to superiority, equivalence, and non-inferiority designs. We argued that the Bayesian approach should be preferred because the conduction of NHSTs in general has, among other things, the disadvantages that evidence in favour of the null hypothesis cannot be quantified (Gallistel, 2009; van Ravenzwaaij et al., 2019; Wagenmakers, 2007; Wagenmakers et al., 2018) and that researchers have to adhere to a predefined sampling plan (e.g., Rouder, 2014; Schönbrodt & Wagenmakers, 2018; Schönbrodt et al., 2017). Moreover, frequentist superiority, equivalence, and non-inferiority tests specifically potentially bear undesired interpretational difficulties (van Ravenzwaaij et al., 2019). Therefore, we provided and explained our baymedr software, written in R (R Core Team, 2019), that allows for the easy computation of Bayes factors for these biomedical designs.

The simulations for the equivalence and non-inferiority designs showed superior performance of the Bayesian compared to the frequentist approach. This difference in performance was less obvious but still consistently observed in the non-inferiority

design. However, an enormous divergence was apparent in the equivalence simulations. In particular, the frequentist TOST equivalence test (Lakens, 2017; Lakens et al., 2018; Schuirmann, 1987) required a very high sample size or a broad equivalence interval to reach sufficient power to detect equivalence. Indeed, this requirement is also reflected in a real example calculation (cf. Eskine, 2013; Moery & Calin-Jageman, 2016; see also Lakens, 2017) that is provided in the documentation of the 'TOSTtwo()' function of the TOSTER software (Lakens, 2017), where a very large equivalence margin of $d = 0.48$ is employed. Another example can be found in a paper by van Dongen et al. (2019), where several researchers were asked to independently analyse the same study according to their preferences. In the study of interest, the incidence rate of birth deficits was compared between women either taking or not taking cetirizine in the first trimester of pregnancy (cf. Weber-Schoendorfer & Schaefer, 2008). Lakens and Hennig chose to conduct an equivalence test, in order to determine whether the groups are practically equivalent. They decided to use an equivalence margin of 10%, representing the smallest effect size of interest (SESOI). Although the two analysts acknowledged that the equivalence margin was chosen more or less arbitrarily (see comments in the R code; https://osf.io/crju5/), we argue that a difference in birth defects of 10% is such an enormous effect that it seems wrong to label this as equivalent in practice.

How Accurate Are Positive Findings?

All of our ROC analyses were based on a balanced proportion of positive (e.g., equivalence and non-inferiority) and negative (e.g., non-equivalence and inferiority) instances. Clearly, however, the proportion of positive and negative instances is not always equal. Take, for instance, the prevalence of a very rare disease. Given that we make a positive prediction (e.g., equivalence or non-inferiority), what is the probability that this prediction is actually true? This probability can be expressed with the positive predictive value (PPV; also called precision or false positive report probability; Wacholder, Chanock, Garcia-Closas, El ghormli, & Rothman, 2004). There are multiple

ways to compute this quantity, one of which is:

$$\frac{TPR \times Prevalence}{TPR \times Prevalence + FPR \times (1 - Prevalence)}, \tag{13}$$

where TPR is the true positive rate, FPR the false positive rate, and $Prevalence$ the rate of truly positive instances.

The PPV depends on the prevalence of truly positive instances, the power of the study, and the amount of evidence obtained (Ioannidis, 2005; Wacholder et al., 2004). Fig 11 shows some benchmarks of PPVs that would be obtained for various true and false positive rates and prevalences. The PPVs are plotted on the y-axis, the true positive rate on the x-axis, and the different lines correspond to different false positive

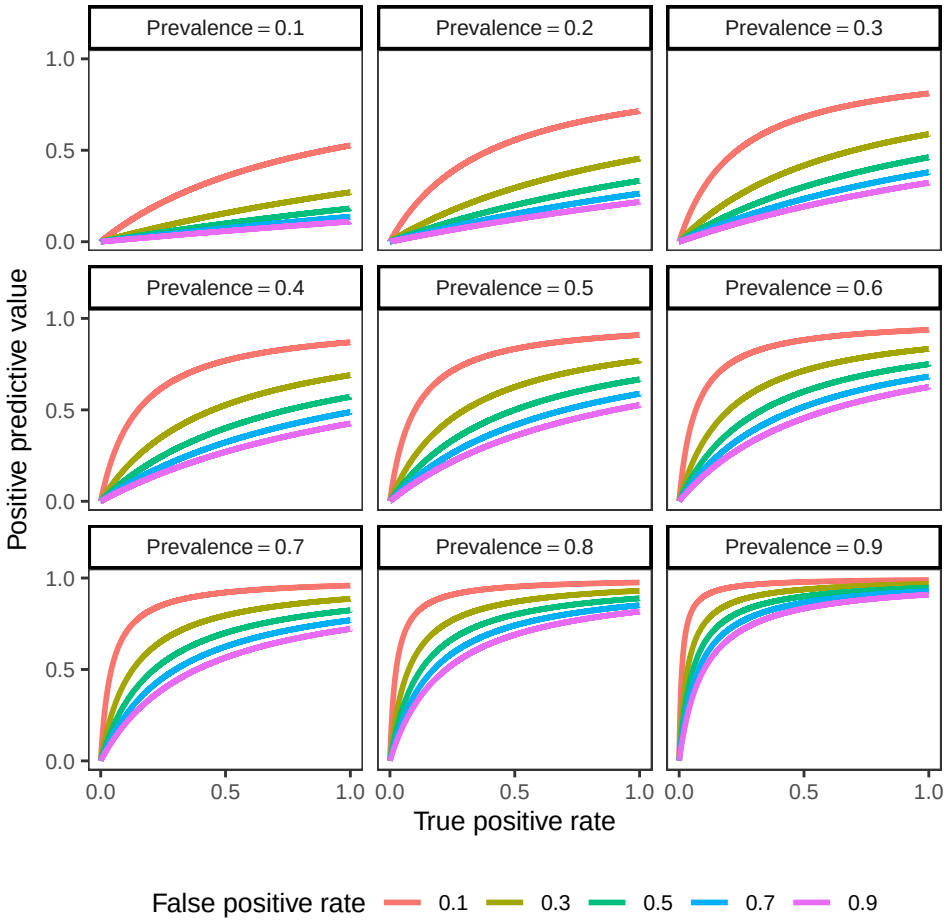

Figure 11. Positive predictive value as a function of the true positive rate, the false positive rate, and the prevalence of truly positive instances.

rates. The panels represent different prevalences. Fig 11 clearly shows that the PPV increases as the prior probability of a truly positive instance (i.e., prevalence) increases. Moreover, the better the test, as expressed with a high true positive rate and low false positive rate, the higher the PPV.

The Prediction Accuracy of Bayesian Inference Is Underestimated

The results of our main simulations show that the Bayesian equivalence and non-inferiority test perform better than the corresponding frequentist counterparts. Still, Bayesian decision making in our simulations was unrealistic because we only defined a threshold for declaring equivalence (or non-inferiority) but no threshold for concluding non-equivalence (or inferiority). Therefore, any Bayes factor that did not exceed this threshold was treated as providing ambiguous evidence. In reality, we would conclude equivalence (or non-inferiority) if the Bayes factor exceeds this threshold (e.g., $BF_{thr} = 10$) and non-equivalence (or inferiority) if the Bayes factor exceeds the reciprocal of that threshold (e.g., $BF_{thr} = 1/10$). Ambiguous evidence would be obtained if the Bayes factor lies between the two thresholds (see Fig 5).

The exploratory analyses of realistic Bayesian decision making showed that the equivalence and non-inferiority tests performed very well. False predictions were only obtained in the cases where very lenient evidence thresholds (e.g., $BF_{thr} = 3$) and small sample sizes are chosen and the true population effect size is close to the equivalence interval boundary (or non-inferiority margin). Still, in many panels, most of the evidence was ambiguous. If that is true, the sceptical reader might ask why we portrait the proper Bayesian equivalence and non-inferiority designs as performing very well. This is because an ambiguous finding is not necessarily a disadvantage. Within the Bayesian framework, we can just sample further cases and monitor the evidence until the desired evidence is reached (Rouder, 2014; Schönbrodt & Wagenmakers, 2018; Schönbrodt et al., 2017). Therefore, it can be argued that the Bayesian equivalence and non-inferiority tests might actually perform better than depicted in the results of the main analyses, which is largely due to the inherent flexibility of the Bayesian approach.

The Prior Distribution

There is a heated and ongoing debate in the literature about how to properly define priors. In particular, it is argued that the specification of the prior distribution is an overly subjective matter and that different prior distributions lead to very different resulting Bayes factors (e.g., Kass & Raftery, 1995; Liu & Aitkin, 2008; Rouder et al., 2009; Sinharay & Stern, 2002; Tendeiro & Kiers, 2019; Vanpaemel, 2010). Further, it is argued that even objective prior distributions (Jeffreys, 1961; Liang et al., 2008; Rouder et al., 2009), that try to remove this subjectivity as good as possible, do not fully solve this problem (e.g., Kruschke & Liddell, 2018a; Tendeiro & Kiers, 2019). Taking, for example, the objective Cauchy prior on effect size, different Cauchy prior scales can result in differing Bayes factors.

Indeed, in certain situations (e.g., with small or moderate sample sizes or unreasonable Cauchy prior scales), the Bayes factor can be sensitive to variations in the prior scale (Rouder et al., 2009). As a point of demonstration, conducting a two-sided Bayesian superiority test in baymedr with a sample size of $n = 100$ in each group, a mean difference of $d = 0.5$, and a standard deviation of $SD = 1$ in both groups, we obtain $BF_{10} \approx 51.6$ with a Cauchy prior scale of $r = 0.5$ and $BF_{10} \approx 9.9$ with a Cauchy prior scale of $r = 5$. However, we believe that we should foremost trust the reason of the researcher, who knows that extremely large effect sizes are not probable and, thus, does not choose outrageously large Cauchy prior scales (cf. Rouder et al., 2009). This already largely narrows down the range of sensible scales. Moreover, it might be possible to utilise knowledge from previous studies (e.g., Dienes, 2011; Lee & Wagenmakers, 2013; Vanpaemel, 2010). If a more or less decent range of potential prior scales is found, a sensitivity/robustness analysis can be conducted (Berger et al., 1994; Du, Edwards, & Zhang, 2019; Kass & Raftery, 1995; Lee & Wagenmakers, 2013). The idea behind a robustness analysis is to calculate several Bayes factors for multiple selections of Cauchy prior scales. Using this approach, the researcher could report the minimum and maximum Bayes factor obtained through the robustness check and openly acknowledge the variability in the results. A robustness check is already

implemented in JASP (JASP Team, 2018) and will be considered in the future for our baymedr software. Lastly, we want to emphasise that even if an unreasonable prior scale was selected, a sceptical reader can recalculate the Bayes factor with his or her own preference for the scale. For this, however, the original researcher must fully disclose the choice of the Cauchy prior scale. Therefore, we stress the importance of transparency in reporting all data-analytic decisions.

Conclusions

Tests of superiority, equivalence, and non-inferiority are important means to compare the effectiveness of medications and treatments in biomedical research. Despite of several limitations, researchers overwhelmingly rely on frequentist inference to analyse the corresponding data for these research designs (Chavalarias et al., 2016). We believe that Bayes factors (Jeffreys, 1939, 1948, 1961; Kass & Raftery, 1995) are an attractive alternative to NHSTs and p-values because they allow researchers to quantify evidence in favour of the null hypothesis (e.g., Gallistel, 2009; van Ravenzwaaij et al., 2019; Wagenmakers, 2007; Wagenmakers et al., 2018) and permit sequential testing and optional stopping (e.g., Rouder, 2014; Schönbrodt & Wagenmakers, 2018; Schönbrodt et al., 2017). In fact, we believe that the possibility for optional stopping and sequential testing has the potential to largely reduce the amount of scarce resources. This is especially important in the field of biomedicine, where clinical trials might be expensive or even harmful for participants.

The baymedr software enables researchers to conduct Bayesian superiority, equivalence, and non-inferiority tests. baymedr is characterised by a very user-friendly implementation, making it convenient for researchers who are not statistical experts. Furthermore, using baymedr, it is possible to calculate Bayes factors based on raw data and summary statistics, which might be valuable for the reanalysis of existing studies.

Using a set of simulations, we demonstrated that Bayesian equivalence and non-inferiority tests consistently display a better classification performance than their conventional frequentist counterparts. These findings hold for various situations (e.g., different effect sizes and samples sizes) that might be encountered in the real world.

Given our compelling results and the theoretical advantages of the Bayesian framework, we therefore encourage researchers to consider the adoption of Bayes factors for the quantification of evidence in equivalence, non-inferiority, and superiority designs.

References

Altman, D. G., & Bland, J. M. (1995, 8). Statistics notes: Absence of evidence is not evidence of absence. *BMJ*, *311*(7003), 485–485. doi: 10.1136/bmj.311.7003.485

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*(6), 423–437. doi: 10.1037/h0020412

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., ... Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. doi: 10.1038/s41562-017-0189-z

Berger, J. O., Moreno, E., Pericchi, L. R., Bayarri, M. J., Bernardo, J. M., Cano, J. A., ... Sivaganesan, S. (1994). An overview of robust Bayesian analysis. *Test*, *3*(1), 5–124. doi: 10.1007/BF02562676

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, *82*(397), 112–122. doi: 10.2307/2289131

Berry, D. A. (2006). Bayesian clinical trials. *Nature Reviews Drug Discovery*, *5*(1), 27–36. doi: 10.1038/nrd1927

Blackwelder, W. C. (1982). "Proving the null hypothesis" in clinical trials. *Controlled Clinical Trials*, *3*(4), 345–353. doi: 10.1016/0197-2456(82)90024-1

Chadwick, D. (1999). Safety and efficacy of vigabatrin and carbamazepine in newly diagnosed epilepsy: A multicentre randomised double-blind study. *The Lancet*, *354*(9172), 13–19. doi: 10.1016/S0140-6736(98)10531-7

Chavalarias, D., Wallach, J. D., Li, A. H. T., & Ioannidis, J. P. A. (2016). Evolution of reporting p values in the biomedical literature, 1990-2015. *Journal of the American Medical Association*, *315*(11), 1141–1148. doi: 10.1001/jama.2016.1952

Christensen, E. (2007). Methodology of superiority vs. equivalence trials and non-inferiority trials. *Journal of Hepatology*, *46*(5), 947–954. doi: 10.1016/j.jhep.2007.02.015

Christensen, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician*, *59*(2), 121–126. doi: 10.1198/000313005X20871

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*(12), 997–1003. doi: 10.1037/0003-066X.49.12.997

Dickey, J. M., & Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, *41*(1), 214–226. doi: 10.1214/aoms/1177697203

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*(3), 274–290. doi: 10.1177/1745691611406920

Du, H., Edwards, M. C., & Zhang, Z. (2019). Bayes factor in one-sample tests of means with a sensitivity analysis: A discussion of separate prior distributions. *Behavior Research Methods*. Retrieved from `http://link.springer.com/10.3758/s13428-019-01262-w` doi: 10.3758/s13428-019-01262-w

Eskine, K. J. (2013). Wholesome foods and wholesome morals?: Organic foods reduce prosocial behavior and harshen moral judgments. *Social Psychological and Personality Science*, *4*(2), 251–254. doi: 10.1177/1948550612447114

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. doi: 10.1016/j.patrec.2005.10.010

Friedman, L. M., Furberg, C. D., DeMets, D. L., Reboussin, D. M., & Granger, C. B. (2010). *Fundamentals of clinical trials* (4th ed.). New York, NY: Springer.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*(2), 439–453. doi: 10.1037/a0015251

Garrett, A. D. (2003). Therapeutic equivalence: Fallacies and falsification. *Statistics in Medicine*, *22*(5), 741–762. doi: 10.1002/sim.1360

Gelman, A. (2013). P values and statistical practice. *Epidemiology*, *24*(1), 69–72. doi: 10.1097/EDE.0b013e31827886f7

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: What you always

wanted to know about significance testing but were afraid to ask. In D. Kaplan (Ed.), *The sage handbook of quantitative methodology for the social sciences* (pp. 391–408). Thousand Oaks, CA: Sage.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. (1995). *Markov chain Monte Carlo in practice*. Boca Raton, FL: Chapman & Hall/CRC.

Goodman, S. N. (1999a). Toward evidence-based medical statistics. 1: The p value fallacy. *Annals of Internal Medicine*, *130*(12), 995–1004. doi: 10.7326/0003-4819-130-12-199906150-00008

Goodman, S. N. (1999b). Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine*, *130*(12), 1005–1013. doi: 10.7326/0003-4819-130-12-199906150-00019

Goodman, S. N. (2008). A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology*, *45*(3), 135–140. doi: 10.1053/j.seminhematol.2008.04.003

Greene, W. L., Concato, J., & Feinstein, A. R. (2000). Claims of equivalence in medical research: Are they supported by the evidence? *Annals of Internal Medicine*, *132*(9), 715–722. doi: 10.7326/0003-4819-132-9-200005020-00006

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2019). Informed Bayesian t -tests. *The American Statistician*, *0*(0), 1–7. doi: 10.1080/00031305.2018.1562983

Hills, R. K. (2017). Non-inferiority trials: No better? No worse? No change? No pain? *British Journal of Haematology*, *176*(6), 883–887. doi: 10.1111/bjh.14504

Hoekstra, R., Monden, R., van Ravenzwaaij, D., & Wagenmakers, E.-J. (2018). Bayesian reanalysis of null results reported in medicine: Strong yet variable evidence for the absence of treatment effects. *PLoS ONE*, *13*(4), e0195474. doi: 10.1371/journal.pone.0195474

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. doi: 10.1371/journal.pmed.0020124

JASP Team. (2018). *JASP (Version 0.9)[Computer software].* Retrieved from `https://jasp-stats.org/`

Jeffreys, H. (1939). *Theory of probability* (1st ed.). Oxford, UK: The Clarendon Press.

Jeffreys, H. (1948). *Theory of probability* (2nd ed.). Oxford, UK: The Clarendon Press.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxfort University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. doi: 10.2307/2291091

Kaul, S., & Diamond, G. A. (2006, 7). Good enough: A primer on the analysis and interpretation of noninferiority trials. *Annals of Internal Medicine*, *145*(1), 62–69. doi: 10.7326/0003-4819-145-1-200607040-00011

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Boston, MA: Academic Press.

Kruschke, J. K., & Liddell, T. M. (2018a). Bayesian data analysis for newcomers. *Psychonomic Bulletin and Review*, *25*(1), 155–177. doi: 10.3758/s13423-017-1272-1

Kruschke, J. K., & Liddell, T. M. (2018b). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin and Review*, *25*(1), 178–206. doi: 10.3758/s13423-016-1221-4

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. doi: 10.1177/1948550617697177

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. doi: 10.1177/2515245918770963

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge, UK: Cambridge University Press. doi: 10.1017/CBO9781139087759

Lesaffre, E. (2008). Superiority, equivalence, and non-inferiority trials. *Bulletin of the NYU Hospital for Joint Diseases*, *66*(2), 150–154.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g

priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*(481), 410–423. doi: 10.1198/016214507000001337

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*(6), 362–375. doi: 10.1016/j.jmp.2008.03.002

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, *5*(6), 161–171. doi: 10.1111/1467-8721.ep11512376

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Meyners, M. (2012). Equivalence tests – A review. *Food Quality and Preference*, *26*(2), 231–245. doi: 10.1016/j.foodqual.2012.05.003

Moery, E., & Calin-Jageman, R. J. (2016). Direct and conceptual replications of Eskine (2013): Organic food exposure has little to no effect on moral judgments and prosocial behavior. *Social Psychological and Personality Science*, *7*(4), 312–319. doi: 10.1177/1948550616639649

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016, 6). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, *72*, 6–18. doi: 10.1016/j.jmp.2015.11.001

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes Factors for Common Designs.* Retrieved from `https://cran.r-project.org/package=BayesFactor`

O'Hagan, A., & Forster, J. (2004). *Kendall's advanced theory of statistics: Vol. 2B. Bayesian inference* (2nd ed.). London, UK: Arnold.

Piaggio, G., Elbourne, D. R., Pocock, S. J., Evans, S. J. W., & Altman, D. G. (2012). Reporting of noninferiority and equivalence randomized trials. *Journal of the American Medical Association*, *308*(24), 2594–2604. doi: 10.1001/jama.2012.87802

R Core Team. (2019). *R: A language and environment for statistical computing.*

Vienna, Austria. Retrieved from https://www.r-project.org/

Ranganathan, P., Pramesh, C. S., & Buyse, M. (2016). Common pitfalls in statistical analysis: The perils of multiple testing. *Perspectives in Clinical Research*, *7*(2), 106–107. doi: 10.4103/2229-3485.179436

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. doi: 10.3758/s13423-014-0595-4

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. doi: 10.3758/PBR.16.2.225

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin and Review*, *25*(1), 128–142. doi: 10.3758/s13423-017-1230-y

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322–339. doi: 10.1037/met0000061

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*(6), 657–680. doi: 10.1007/BF01068419

Sinharay, S., & Stern, H. S. (2002). On the sensitivity of Bayes factors to the prior distributions. *The American Statistician*, *56*(3), 196–201. doi: 10.1198/000313002137

Stanislaw, H., & Todorov, N. (1999, 3). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149. doi: 10.3758/BF03207704

Tendeiro, J. N., & Kiers, H. A. (2019). A review of issues about null hypothesis bayesian testing. *Psychological Methods*. doi: http://dx.doi.org/10.1037/met0000221

The jamovi project. (2019). *jamovi (Version 0.9)[Computer software].* Retrieved from https://www.jamovi.org

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, *22*(2), 217–239. doi: 10.1037/met0000100.supp

Van de Werf, F., Adgey, J., Ardissino, D., Armstrong, P. W., Aylward, P., Barbash, G., ... White, H. (1999). Single-bolus tenecteplase compared with front-loaded alteplase in acute myocardial infarction: the ASSENT-2 double-blind randomised trial. *The Lancet*, *354*(9180), 716–722. doi: 10.1016/S0140-6736(99)07403-6

van Dongen, N. N. N., van Doorn, J. B., Gronau, Q. F., van Ravenzwaaij, D., Hoekstra, R., Haucke, M. N., ... Wagenmakers, E.-J. (2019). Multiple perspectives on inference for two simple statistical scenarios. *The American Statistician*, *73*(sup1), 328–339. doi: 10.1080/00031305.2019.1565553

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*(6), 491–498. doi: 10.1016/j.jmp.2010.07.003

van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to Markov chain Monte–Carlo sampling. *Psychonomic Bulletin & Review*, *25*(1), 143–154. doi: 10.3758/s13423-016-1015-8

van Ravenzwaaij, D., & Ioannidis, J. P. A. (2017). A simulation study of the strength of evidence in the recommendation of medications based on two trials with statistically significant results. *PLoS ONE*, *12*(3), e0173184. doi: 10.1371/journal.pone.0173184

van Ravenzwaaij, D., & Ioannidis, J. P. A. (2018). *True and false positive rates for different criteria of evaluating statistical evidence from clinical trials.* Retrieved from `https://doi.org/10.31222/osf.io/kcz3y`

van Ravenzwaaij, D., Monden, R., Tendeiro, J. N., & Ioannidis, J. P. A. (2019). Bayes factors for superiority, non-inferiority, and equivalence designs. *BMC Medical Research Methodology*, *19*(1), 71. doi: 10.1186/s12874-019-0699-7

Wacholder, S., Chanock, S., Garcia-Closas, M., El ghormli, L., & Rothman, N. (2004). Assessing the probability that a positive report is false: An approach for

molecular epidemiology studies. *Journal of the National Cancer Institute*, *96*(6), 434–442. doi: 10.1093/jnci/djh075

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. doi: 10.3758/BF03194105

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology*, *60*(3), 158–189. doi: 10.1016/j.cogpsych.2009.12.001

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57. doi: 10.3758/s13423-017-1343-3

Walker, E., & Nowacki, A. S. (2011). Understanding equivalence and noninferiority testing. *Journal of General Internal Medicine*, *26*(2), 192–196. doi: 10.1007/s11606-010-1513-8

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. doi: 10.1080/00031305.2016.1154108

Weber-Schoendorfer, C., & Schaefer, C. (2008). The safety of cetirizine during pregnancy: A prospective observational cohort study. *Reproductive Toxicology*, *26*(1), 19–23. doi: 10.1016/j.reprotox.2008.05.053

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparson using 855 t tests. *Perspectives on Psychological Science*, *6*(3), 291–298. doi: 10.1177/1745691611406923

Wickham, H. (2019). *Advanced R* (2nd ed.). Boca Raton, FL: CRC Press.

Winkler, R. L. (2001). Why Bayesian analysis hasn't caught on in healthcare decision making. *International Journal of Technology Assessment in Health Care*, *17*(1), 56–66. doi: 10.1017/S026646230110406X

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A

fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, *39*(4), 561–577.