medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



# Bayes Factors for Two-group Comparisons in Cox Regression

Maximilian Linde[1], Jorge N. Tendeiro[2], and Don van Ravenzwaaij[1]

[1]Unit of Psychometrics and Statistics, Department of Psychology, Faculty of Behavioural and Social Sciences, University of Groningen, Groningen, The Netherlands

[2]Office of Research and Academia-Government-Community Collaboration, Education and Research Center for Artificial Intelligence and Data Innovation, Hiroshima University, Hiroshima, Japan

## Author Note

Correspondence concerning this article should be addressed to: Maximilian Linde, University of Groningen, Department of Psychology, Grote Kruisstraat 2/1, Heymans Building, Room 217, 9712 TS Groningen, The Netherlands, Phone: (+31) 50 363 2702, E-mail: m.linde@rug.nl.

Maximilian Linde https://orcid.org/0000-0001-8421-090X

Jorge N. Tendeiro https://orcid.org/0000-0003-1660-3642

Don van Ravenzwaaij https://orcid.org/0000-0002-5030-4091

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



## Abstract

The use of Cox proportional hazards regression to analyze time-to-event data is ubiquitous in biomedical research. Typically, the frequentist framework is used to draw conclusions about whether hazards are different between patients in an experimental and a control condition. We offer a procedure to calculate Bayes factors for simple Cox models, both for the scenario where the full data is available and for the scenario where only summary statistics are available. The procedure is implemented in our "baymedr" R package. The usage of Bayes factors remedies some shortcomings of frequentist inference and has the potential to save scarce resources.

*Keywords:* Bayes factor, Cox proportional hazards regression, particle swarm optimization, survival, time-to-event data

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



## Bayes Factors for Two-group Comparisons in Cox Regression

The biomedical literature is filled with studies in which two conditions are compared on some kind of outcome measure. A common example is a clinical trial in which the goal is to determine the efficacy of a therapeutic agent over a placebo or an already existing medication (e.g., Christensen, 2007; Friedman et al., 2010; Senn, 2008). The outcome measure can be continuous; an example would be symptom severity. Alternatively, the outcome measure can be dichotomous, as in studies that examine the mortality of patients following a medical procedure (see, e.g., Dean et al., 2001; Pearse et al., 2012). Sometimes, not only the sheer absence or presence of some event is relevant, but also the time until that event happens, which is often called the survival or failure time (Harrell, 2015). For instance, in order to judge the effectiveness of some form of oncological treatment, it is of interest to know how long terminally ill cancer patients survive after receiving the treatment, and at which time there is an increased or decreased risk of death.

Time-to-event data are typically analyzed using *survival analysis* (see Bradburn et al., 2003a, 2003b; Clark et al., 2003a, 2003b; Collett, 2015; Harrell, 2015; Hosmer et al., 2008, for excellent overviews). Usually, researchers use the frequentist statistical framework for survival analyses. This, however, has several disadvantages: First, it is impossible to quantify evidence *in favor* of the null hypothesis (e.g., Rouder et al., 2009) of equal survival between conditions. The reason for that is that a non-significant finding can occur due to low power or a truly absent effect; the two possibilities cannot be disentangled (Bakan, 1966; Keysers et al., 2020; van Ravenzwaaij et al., 2019). Second, stopping data collection based on interim results (e.g, $p$-value already reached threshold or $p$-value did not yet reach threshold) is highly problematic because it increases the probability of having a false positive result (Armitage et al., 1969; Rouder, 2014; Tendeiro et al., 2022).

We offer a procedure and easy-to-use implementation for hypothesis testing for survival analysis in the Bayesian framework. Specifically, we focus on *Cox proportional hazards regression* (henceforth called either Cox regression or Cox model; Cox, 1972). This

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



allows directly contrasting the evidence for the null hypothesis that there is no effect with an alternative hypothesis that operationalizes that there is some effect; it also allows monitoring results and stopping data collection at will. Moreover, to the best of our knowledge, so far Cox regression can only be conducted when the full data is available. Oftentimes, however, it is relevant to reanalyze studies based on summary statistics reported in articles (e.g., in replications and meta analyses; e.g., Field & Gillett, 2010; Sutton & Abrams, 2001). We describe how data can be simulated based on summary statistics and how these simulated data can be used to subsequently conduct Bayesian hypothesis testing for Cox regression. We implemented both the process of data simulation and the process of Bayesian hypothesis testing in a software package (the "baymedr" R package; Linde et al., 2022) that can be used by a wide audience of researchers, as we will illustrate.

The remainder of the article is organized as follows. First, we give an introduction to survival analysis in general and Cox regression in particular. Second, we explain how Bayes factors can be computed and interpreted and apply this to the special case of Cox regression. Third, we showcase how our "baymedr" software can be used to calculate a Bayes factor when the full data is available. Fourth, we describe our procedure for the scenario where only summary statistics are available. Specifically, we describe how survival data can be simulated from summary statistics, we tune parameters for the data simulation process, and demonstrate how "baymedr" can be used to calculate a distribution of Bayes factors for multiple simulated data sets. Fifth, we compare the performance (in terms of bias and variance) of our approach to an approximation approach advocated by Bartoš and Wagenmakers (2022).

## Survival Analysis

The time until the event of interest occurs is often called *survival time* or failure time. Typically, the survival time is only known for some participants. Reasons for that

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



include the study ending before the event of interest is observed, participants withdrawing from the study, or failure to follow up with participants (e.g., a participant might have moved without notifying the researchers). In these cases, participants' survival time is censored on the right. That is, the exact survival time is not known; what is known, however, is that the survival time exceeds the time at the last follow-up (see, e.g., Harrell, 2015; Hosmer et al., 2008; Klein & Moeschberger, 1997; Leung et al., 1997, for more thorough treatments of censoring). For example, assume that a study is scheduled for a time period of seven years. A participant enters the study two years after the study has started. At the end of the study, the event of interest still has not occurred for this participant. In that case, the observation of five years is censored, which means that it is only known that the survival time of that participant exceeds five years. One advantage of survival analysis is that it can handle this kind of incomplete data very well.

We denote the survival time as $T$. The cumulative distribution function of $T$ gives the probability that a participant has a survival time that is equal or less than some particular time $t$:

$$F(t) = P(T \leq t). \tag{1}$$

Oftentimes, the probability that a participant is still alive after a particular time $t$ is more interesting. This is given by the *survival function*:

$$S(t) = P(T > t) = 1 - F(t). \tag{2}$$

Lastly, it is also informative to examine time periods of increased and decreased risk of failure. This is not immediately apparent in the survival function. The *hazard function* displays the instantaneous risk that the event happens in a narrow interval around a particular time $t$ for participants who already survived until time $t$ (cf. Clark et al., 2003a):

$$\lambda(t) = \lim_{u \to 0} \frac{P(t < T \leq t + u \mid T > t)}{u}. \tag{3}$$

This is equivalent to:

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{\partial \log S(t)}{\partial t}, \tag{4}$$

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



with $f(t)$ being the probability density function of $T$ (see, Harrell, 2015, pp. 404–405, for a derivation of this equality).

Various kinds of survival analyses exist for estimating $S(t)$ and $\lambda(t)$. These approaches can be grouped into one of three categories: non-parametric, parametric, and semi-parametric approaches. The non-parametric approach does not make any assumptions about the survival and hazard functions. The most prominent non-parametric approach is the Kaplan-Meier product-limit estimator (Kaplan & Meier, 1958), which is often considered as a first descriptive step (Harrell, 2015). In contrast to the non-parametric approach, the parametric approach makes some assumptions on the survival function. Specifically, the survival function is assumed to belong to a predetermined distribution family, whose parameters need to be estimated in a way that fits the data best. Common distribution families for the survival function include the Exponential, Weibull, and Gompertz distributions. Finally, the semi-parametric Cox regression (Cox, 1972) is used most commonly across all types of survival analyses. On the one hand, it is parametric because it assumes a multiplicative effect of the predictors on the hazard function (i.e., the assumption of proportional hazards); on the other hand, it is non-parametric because it does not impose any particular form on the hazard function. The focus of this paper lies exclusively on Cox regression.

## Cox Regression

In Cox regression, the data for each participant $i$, with $i \in \{1, \ldots, n\}$, consists of the observed response $Y_i$ and an event indicator $\delta_i$, designating whether the event of interest occurred (1) or not (0). In addition, each participant has a censoring time $C_i$. Each participant has a survival time $T_i$ that is known in case this observation is uncensored (i.e., $\delta_i = 1$) or only known to be greater than some value in case this observation is censored (i.e., $\delta_i = 0$). The observed response $Y_i$ is:

$$Y_i = \min(T_i, C_i). \tag{5}$$

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



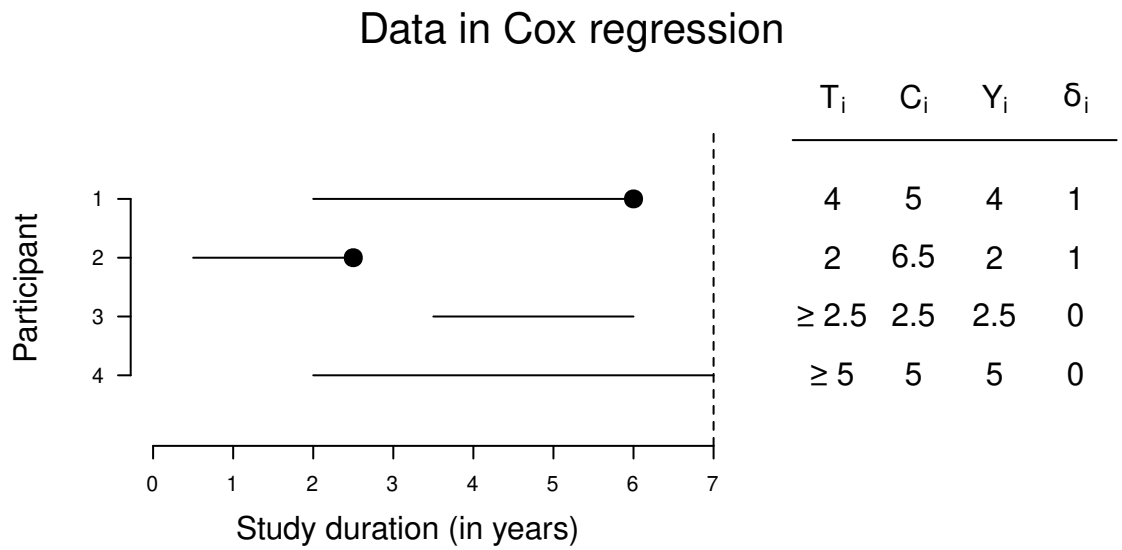


**Figure 1**

*Example of data in Cox regression with four participants. The horizontal lines show the individual observed responses (i.e., survival/censoring times). The solid dots indicate that the event of interest was observed. The vertical dashed line demarcates the end of the study. $T$ is the survival time, $C$ is the censoring time, $Y$ is the observed response, and $\delta$ is the event indicator.*

Thus, the relevant information for the $i$-th participant is fully contained in the combination of $Y_i$ and $\delta_i$. This is visualized in Figure 1.

In the following, we will assume that we have one independent variable $x$ that is dichotomous, indicating membership to one of two conditions. Therefore, each participant $i$ has a value on the independent variable, say $x_i$. Importantly, we will assume that $x_i = 0$ refers to the control condition ($c$) and $x_i = 1$ to the experimental condition ($e$). We will refer to the combination of $Y$, $\delta$, and $x$ as the data $D$ for a generic participant. The Cox model (Cox, 1972) is expressed as:

$$\lambda\left(t \mid x\right) = \lambda\left(t\right) e^{x\beta}, \tag{6}$$

where $\beta$ is the parameter we aim to estimate. $\lambda\left(t \mid x\right)$ is the hazard function, $\lambda\left(t\right)$ is the

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



baseline hazard function indicating how $\lambda(t \mid x)$ changes as a function of $t$, and $e^{x\beta}$ is the relative hazard function characterizing how $\lambda(t \mid x)$ changes as a function of $x$ (Hosmer et al., 2008). Assuming that $x_i \in \{0,\ 1\}$, the hazard ratio is:

$$\mathrm{HR} = \frac{\lambda(t \mid x = 1)}{\lambda(t \mid x = 0)} = \frac{\lambda(t)\, e^{\beta}}{\lambda(t)} = e^{\beta}. \tag{7}$$

The estimation of $\beta$ is based most often on maximum likelihood estimation. Commonly, either the original Cox partial likelihood (Cox, 1972), Breslow's approximation to the true partial likelihood (Breslow, 1974), or Efron's approximation to the true partial likelihood (Efron, 1977) are used (see, e.g., Therneau & Grambsch, 2000, for a more detailed discussion). It is a partial instead of a true likelihood because it does not take the actual responses but only their rank ordering into account (Collett, 2015). We will exclusively use Efron's method because it usually is most accurate, can handle tied survival times best, and is the default in several software packages; examples are the "survival" R package (Therneau, 2021; Therneau & Grambsch, 2000) and the "rms" R package (Harrell, 2022).

Let $t_j$ represent the $k$ (where $k \leq n$) unique ordered survival times, with $j \in \{1,\ \ldots,\ k\}$. Then, for a generic independent variable, Efron's approximation of the true log partial likelihood is defined as (cf. Harrell, 2015, p. 477):

$$\log \mathcal{L}(\beta) = \sum_{j=1}^{k} \left[ \left( \sum_{l \in N_j} x_l \right) \beta - \sum_{b=1}^{n_j} \log \left[ \left( \sum_{h \in S_j} e^{x_h \beta} \right) - \frac{b-1}{n_j} \left( \sum_{l \in N_j} e^{x_l \beta} \right) \right] \right], \tag{8}$$

where $N_j$ is the set of indices for cases failing at $t_j$, $n_j$ is the number of cases failing at $t_j$, and $S_j$ is the set of indices for cases having an observed response of at least $t_j$.

Once the model is estimated, a confidence interval for $\beta$ or HR can be calculated. Alternatively, null hypothesis significance testing (NHST) in the form of a Wald test (for instance) is conducted, with the null hypothesis:

$$\mathcal{H}_0 \colon \beta = \beta_0 \tag{9}$$

and the alternative hypothesis either being two-sided:

$$\mathcal{H}_1 \colon \beta \neq \beta_0 \tag{10}$$

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



or one-sided:

$$\mathcal{H}_1\colon \beta < \beta_0 \qquad \text{or} \qquad \mathcal{H}_1\colon \beta > \beta_0, \tag{11}$$

with $\beta_0$ being the null value. Rejection of $\mathcal{H}_0$ is warranted when the resulting $p$-value is smaller than a predefined significance level (when $p < \alpha$); when $p \geq \alpha$ nothing can be concluded.

In the next section, we turn our attention to Bayes factors. We describe the motivation and theory behind Bayes factors, how they are interpreted, and how they can be calculated for Cox models.

## The Bayes Factor

The use of NHST in biomedical research is ubiquitous (Chavalarias et al., 2016) even though it has been criticized repeatedly (e.g., Berger & Delampady, 1987; Berger & Sellke, 1987; Cohen, 1994; Dienes, 2011; Gigerenzer, 2004; Goodman, 1999a, 1999b, 2008; McShane et al., 2019; van Ravenzwaaij & Ioannidis, 2017; Wagenmakers, 2007; Wagenmakers et al., 2018; Wasserstein & Lazar, 2016). Null hypothesis Bayesian testing (NHBT) is an alternative to NHST that has some practical advantages. For instance, in contrast to NHST, NHBT allows quantifying the relative evidence in favor of $\mathcal{H}_0$ (Rouder et al., 2009; van Ravenzwaaij et al., 2019; Wagenmakers, 2007; Wagenmakers et al., 2018). That way, the evidence for (or against) $\mathcal{H}_0$ and $\mathcal{H}_1$ can be directly compared. This possibility is important because it enables researchers to investigate whether a therapeutic agent is *not* working. Moreover, in contrast to NHST, NHBT enables researchers to monitor the data during data collection and stop or continue data collection as needed (Rouder, 2014; Sanborn & Hills, 2014; Schönbrodt & Wagenmakers, 2018; Schönbrodt et al., 2017; Tendeiro et al., 2022). This might have the implication that fewer resources are wasted because neither too many nor too few cases are sampled (Chalmers & Glasziou, 2009). In addition, the results of NHBT are easy to interpret and are arguably more in line with researchers' questions compared to NHST.

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



The most common vehicle of NHBT is the *Bayes factor* (Jeffreys, 1939, 1948, 1961; Kass & Raftery, 1995), which quantifies the relative probabilities of the data under $\mathcal{H}_0$ and $\mathcal{H}_1$. For example, a Bayes factor of $\mathrm{BF}_{10} = 1$ (the subscript indicates a comparison is made between the evidence for $\mathcal{H}_1$ relative to $\mathcal{H}_0$) indicates a perfect balance between $\mathcal{H}_0$ and $\mathcal{H}_1$, given the choice of prior and model; the data are equally probable under both hypotheses. In contrast, $\mathrm{BF}_{10} = 10$ suggests that the data are 10 times more likely under $\mathcal{H}_1$ compared to $\mathcal{H}_0$, given the choice of prior and model. Lastly, $\mathrm{BF}_{10} = 0.1$ indicates that the data are $\mathrm{BF}_{01} = 1/\mathrm{BF}_{10} = 10$ times more likely under $\mathcal{H}_0$ compared to $\mathcal{H}_1$, given the choice of prior and model.

The Bayes factor follows from applying Bayes' rule to both $\mathcal{H}_1$ and $\mathcal{H}_0$ using the available data $D$, while assuming that $\mathcal{H}_1$ and $\mathcal{H}_0$ are the two only models of interest:

$$\underbrace{\frac{P\left(\mathcal{H}_1 \mid D\right)}{P\left(\mathcal{H}_0 \mid D\right)}}_{\text{Posterior odds}} = \underbrace{\frac{P\left(D \mid \mathcal{H}_1\right)}{P\left(D \mid \mathcal{H}_0\right)}}_{\text{Bayes factor, } \mathrm{BF}_{10}} \times \underbrace{\frac{P\left(\mathcal{H}_1\right)}{P\left(\mathcal{H}_0\right)}}_{\text{Prior odds}}. \tag{12}$$

The prior odds reflect one's initial beliefs about the probabilities of $\mathcal{H}_1$ and $\mathcal{H}_0$, the Bayes factor quantifies the relative probabilities of the data under $\mathcal{H}_1$ and $\mathcal{H}_0$, and the posterior odds reflect the relative probabilities of $\mathcal{H}_1$ and $\mathcal{H}_0$ after having observed the data. It can be seen in Equation 12 that the Bayes factor is independent of the prior odds. Therefore, when people hold different beliefs about the prior odds, they obtain different posterior odds but the Bayes factor remains the same.

The numerator of the Bayes factor in Equation 12 is calculated by integrating the product of the prior distribution for the parameter of interest under the alternative hypothesis and the likelihood function. In the case of Cox regression with one dichotomous independent variable, the parameter of interest is $\beta$. Consequently, the numerator of the Bayes factor in Equation 12 is:

$$P\left(D \mid \mathcal{H}_1\right) = \int_{\beta \in \Omega_1} f\left(D \mid \beta\right) f\left(\beta\right) \mathrm{d}\beta, \tag{13}$$

where $\Omega_1$ is the range of $\beta$ parameter values under $\mathcal{H}_1$. $P\left(D \mid \mathcal{H}_1\right)$ is a marginal likelihood because the $\beta$ parameter is integrated out. One can think of $P\left(D \mid \mathcal{H}_1\right)$ as a weighted

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



average of the likelihood $f(D \mid \beta)$, with weights given by the prior $f(\beta)$ (e.g., Kruschke, 2015; Tendeiro & Kiers, 2019). If a point $\mathcal{H}_0$ is used, the denominator of the Bayes factor in Equation 12 is simply the density of the likelihood evaluated at the null value $\beta_0$:

$$P(D \mid \mathcal{H}_0) = f(D \mid \beta = \beta_0). \tag{14}$$

Even though the Bayes factor is independent of the prior odds, it is sensitive to the choice of prior for $\beta$ (e.g., Gallistel, 2009; Kass & Raftery, 1995; Sinharay & Stern, 2002; Vanpaemel, 2010).

The Bayes factor is calculated the same way for either the full data or for data simulated based on summary statistics. The Bayes factor is:

$$\mathrm{BF}_{10} = \frac{\int_{\beta \in \Omega_1} f(D \mid \beta) f(\beta) \, \mathrm{d}\beta}{f(D \mid \beta = \beta_0)}, \tag{15}$$

where $D$ can refer either to the full data or the simulated data. Importantly, for our application of Cox regression, $f(D \mid \beta)$ is equivalent to the natural exponent of Efron's approximation to the true log partial likelihood (see Equation 8).

A computational challenge is that some parts of Equation 8 can yield such large or small values that they cannot easily be represented by a computer, in which case we encounter overflow and underflow of floating-point numbers, respectively (see, e.g., Goldberg, 1991, for a thorough treatment of floating-point arithmetic). In general, when R encounters overflow, it represents the number as infinity and when R encounters underflow, the resulting value is 0. The parts of Equation 8 that are at risk of under- and overflow are $\sum_{h \in S_j} e^{x_h \beta}$ and $\sum_{l \in N_j} e^{x_l \beta}$. Specifically, these parts are at risk when $\beta$ is large in magnitude.

To overcome this problem, we adapted Equation 8 by adding a well-chosen number $z_j$ to the exponents of the problematic parts:

$$\log \mathcal{L}(\beta) = \sum_{j=1}^{k} \left[ \left( \sum_{l \in N_j} x_l \right) \beta - \left( \left( \sum_{b=1}^{n_j} \log \left[ \left( \sum_{h \in S_j} e^{x_h \beta + z_j} \right) - \frac{b-1}{n_j} \left( \sum_{l \in N_j} e^{x_l \beta + z_j} \right) \right] \right) - n_j z_j \right) \right]. \tag{16}$$

To be clear, Equations 8 and 16 are equivalent for a dichotomous independent variable that is coded with 0 and 1; they are just parameterized differently. In general, for each $j$, let $x_{S_j}$

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



be a vector only containing those $x$-values that have indices included in the set $S_j$. Then, $z_j$ is chosen such that:

$$z_j = \begin{cases} -\beta, & \text{if } \left(\exists x_{S_j}\beta \,.\, x_{S_j}\beta > 0\right) \cup \left(\forall x_{S_j}\beta \,.\, x_{S_j}\beta < 0\right) \\ & \text{(i.e., if any element of } x_{S_j}\beta \text{ is larger than 0} \\ & \text{or all elements of } x_{S_j}\beta \text{ are smaller than 0)} \\ 0, & \text{otherwise.} \end{cases} \tag{17}$$

In other words, for a dichotomous independent variable that is coded with 0 and 1, $z_j = -\beta$ if $x_{S_j}$ contains *any* cases of the experimental group and $\beta > 0$ or when $x_{S_j}$ contains *only* cases of the experimental group and $\beta < 0$. $z_j$ serves the purpose of scaling the powers in case they would result in under- or overflow. Since this scaling is done $n_j$ times, this must be compensated by subtracting $n_j z_j$ at the end of Equation 16.

Still, the calculation of the Bayes factor (see Equation 15) is often not possible because using the natural exponent of Equation 16 (which is $f(D \mid \beta)$) yields very small values for certain ranges of $\beta$ values, so that there is again the problem of underflow. To overcome this obstacle, we apply another transformation (which is inspired by Cook, 2012). First, the maximum $m$ of the sum of the log likelihood and the log prior for $\beta$ must be found:

$$m = \max\left(\log\left[f(D \mid \beta)\, f(\beta)\right]\right) = \max\left(\log\left[f(D \mid \beta)\right] + \log\left[f(\beta)\right]\right). \tag{18}$$

We achieve this by means of numerical optimization using the "Brent" method of the "optim()" function in R (Brent, 1973). $m$ is used as a scaling factor that avoids underflow. Then an integral $I$ is calculated:

$$I = \int_{\beta \in \Omega_1} \exp\left(\log\left[f(D \mid \beta)\right] + \log\left[f(\beta)\right] - m\right) \mathrm{d}\beta. \tag{19}$$

The integral is determined by means of Gaussian quadrature using the "integrate()" function in R. Finally, the log marginal likelihood is a combination of $m$ and $I$:

$$\log\left[P(D \mid \mathcal{H}_1)\right] = \log\left[\int_{\beta \in \Omega_1} f(D \mid \beta)\, f(\beta)\, \mathrm{d}\beta\right] = m + \log\left[I\right]. \tag{20}$$

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



Having this, the Bayes factor can be calculated:

$$\mathrm{BF}_{10} = \exp\left(\log\left[P\left(D \mid \mathcal{H}_1\right)\right] - \log\left[P\left(D \mid \mathcal{H}_0\right)\right]\right). \tag{21}$$

## The Need for Methods for Full Data and Summary Statistics

In the previous two sections, we learned about Cox regression with a dichotomous independent variable and how its regression coefficient ($\beta$ or HR) is estimated. Further, we explained the mathematical details of how to calculate a Bayes factor in general and for a Cox regression specifically.

The calculation of Bayes factors can be challenging for applied researchers who do not have a firm background in Bayesian statistics and programming. Fortunately, multiple software packages are available that allow calculating Bayes factors for various research designs. Examples are the R packages "BayesFactor" (Morey & Rouder, 2018) and "baymedr" (Linde et al., 2022), and point-and-click software like "JASP" (JASP Team, 2022). Moreover, for Bayesian parametric survival analysis, there is the "RoBSA" R package (Bartoš, 2022). However, to the best of our knowledge, there is yet no software implementation that allows calculating Bayes factors for Cox models. In addition to a module for calculating Bayes factors with the full data set at hand, we also include a module that provides a highly accurate approximation of the Bayes factor when only summary statistics are available, for instance in the scenario where one reanalyses the results of a published study with only quantities published in the paper available.

In the next two sections, we showcase how researchers can use our "baymedr" R package (Linde et al., 2022) to calculate Bayes factors for Cox models. In the first section, we focus on the situation where the full data is available. Subsequently, the second section focuses on the situation where only summary statistics are available. In that case, data must be simulated from the summary statistics; we explain how this is done, we tune data simulation parameters, and examine the bias and variance of the simulated Bayes factors.

The files with code for all computations can be found online (available at

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



https://osf.io/37ut2/).

## Calculating a Bayes Factor from the Full Data

We applied our approach for calculating Bayes factors for Cox models to an empirical data set. The source of the data set is a study that is reported in Beigel et al. (2020). The goal of this double-blind, randomized, placebo-controlled trial by Beigel et al. (2020) was to determine the effectiveness of a therapeutic agent called Remdesivir for the treatment of the coronavirus disease 2019 (Covid-19). Participants were $n = 1,062$ adults who were admitted to hospital due to Covid-19 infection. Participants were randomly assigned to a placebo condition ($n_c = 521$) or a Remdesivir condition ($n_e = 541$). The primary outcome was the time until recovery, which was conceptualized as patients being dismissed from the hospital or as patients remaining in hospital solely for the purpose of infection control.

Beigel et al. (2020) conducted a Cox regression to investigate the time until recovery. The authors used group membership (placebo vs. Remdesivir) as the independent variable and stratified by actual disease severity (severe disease vs. mild-moderate disease). In their supplementary material called "Protocol", it is mentioned that a superiority alternative hypothesis is used (i.e., a one-sided hypothesis with HR $> 1$) with a two-sided significance level of $\alpha = .05$. More specifically, the authors hypothesized that patients receiving Remdesivir recover quicker than patients receiving a placebo, which corresponds to patients receiving Remdesivir having higher hazards than patients receiving a placebo. In contrast to our approach, the authors seem to have used Breslow's instead of Efron's approximation to the true partial likelihood. The authors conclude that "[p]atients in the [R]emdesivir group had a shorter time to recovery than patients in the placebo group" (Beigel et al., 2020, p. 1816) and report a hazard ratio (i.e., recovery rate ratio) of HR $= 1.29$ together with a confidence interval of 95% CI $= [1.12,\ 1.49]$ (see Table 2 of Beigel et al., 2020).

Our reanalysis of Beigel et al. (2020) omitted the stratification by actual disease

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



severity. Furthermore, we used Efron's instead of Breslow's approximation to the true partial likelihood. The reason for these two deviations is that we have not implemented them in our "baymedr" R package. The resulting HR and the corresponding confidence interval are very close to Beigel et al. (2020): HR = 1.312, 95% CI = [1.136, 1.514].

For the reanalysis we used our "baymedr" software, which can be downloaded and installed from GitHub using the "devtools" package (Wickham et al., 2019) and loaded by typing the following into the R console:

```
devtools::install_github("maxlinde/baymedr") # Download baymedr
library("baymedr")                           # Attach baymedr
```

Using "baymedr", we can calculate a Bayes factor for the full data in Beigel et al. (2020). The data must have the survival time, the event indicator, and the independent variable in that order as columns (see columns 2, 3, and 4 of Table 1). We used a truncated (because we have a one-sided alternative hypothesis) Normal prior for $\beta$ with a mean of $\mu = 0$ and a standard deviation of $\sigma = 1$:

```
act_bf <- coxph_bf(            # See ?coxph_bf for details
    data        = act_data,    # Object containing the data
    null_value  = 0,           # H0 value
    alternative = "one.sided", # H1 type (one- or two-sided)
    direction   = "high",      # H1 direction (low or high)
    prior_mean  = 0,           # Beta prior mean
    prior_sd    = 1            # Beta prior SD
)
```

The result is object `act_bf`, containing all the relevant information of the Bayes factor analysis. Moreover, typing `act_bf` into the R console results in the printing of a concise summary that includes the type of analysis, the null and alternative hypotheses, the choice of prior, and the obtained Bayes factor:

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



```
******************************
Cox proportional hazards analysis
---------------------------------
H0:                beta == 0
H+:                beta > 0
Normal prior:      Mean = 0.000
                   SD = 1.000


    BF+0 = 134.401
******************************
```

The obtained Bayes factor, $\mathrm{BF}_{+0} = 134.401$, suggests that the data are 134 times more likely to have been generated under the alternative hypothesis compared to the null hypothesis, given the choice of prior and model. Thus, according to approximate Bayes factor thresholds proposed by Kass and Raftery (1995), we found strong evidence for the hypothesis that patients receiving Remdesivir recover quicker than patients receiving a placebo.

So far, we have dealt with situations where the full data set is available. In practice, one may be interested in revisiting published results from an existing study, without having access to the complete data set. In what follows, we will demonstrate how our method can give a very accurate approximation to the Bayes factor when only published summary statistics are available.

## Calculating Bayes factors from Summary Statistics

When researchers conduct their own studies, they have the full data at hand and can use "baymedr" to calculate a Bayes factor. In other situations, however, the full data unfortunately might not be available. For example, when conducting a reanalysis of study findings, some researchers might not be allowed (e.g., for ethical reasons) or willing to

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



share the full data. In these situations, it is paramount to be able to use summary statistics reported in the original manuscript to calculate a Bayes factor.

When the full data is not available, an attractive alternative is to simulate data within the constraints of summary statistics that are known. When following such an approach, it is imperative to demonstrate that the simulated data are sufficiently constrained by the available summary statistics. Our approach for data simulation is based on summary statistics that are commonly reported in scientific articles. Initially, we considered the following candidates:

- Sample sizes within each condition, $n_c$ and $n_e$, respectively,
- Number of events within each condition, $v_c$ and $v_e$, respectively,
- Maximum observed or maximum possible survival time, $t_{\max}$,
- Kaplan-Meier (KM; Kaplan & Meier, 1958) median survival times within each condition, with corresponding confidence intervals, $\mathrm{KM}_c$, $\mathrm{CI}\left(\mathrm{KM}_c\right)_{LB}$, $\mathrm{CI}\left(\mathrm{KM}_c\right)_{UB}$, and $\mathrm{KM}_e$, $\mathrm{CI}\left(\mathrm{KM}_e\right)_{LB}$, $\mathrm{CI}\left(\mathrm{KM}_e\right)_{UB}$, respectively,
- Hazard ratio obtained from a Cox model (Cox, 1972), with corresponding confidence interval, HR, $\mathrm{CI}\left(\mathrm{HR}\right)_{LB}$, and $\mathrm{CI}\left(\mathrm{HR}\right)_{UB}$.

The first step for simulating data is to sample $n_c + n_e$ responses $Y$ drawn from a Uniform distribution:

$$Y \sim \mathrm{Uniform}\left(1,\ t_{max}\right). \tag{22}$$

The choice of a Uniform distribution is arbitrary. Any other probability function would be equally suitable; even sampling $n_c + n_e$ times the same value would suffice. These generated responses are paired with an event indicator $\delta$:

$$\delta = [0,\ 1,\ 0,\ 1]\,, \tag{23}$$

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



whose elements are repeated $n_c - v_c$, $v_c$, $n_e - v_e$, and $v_e$ times, respectively. Lastly, the independent variable $x$ is added:

$$x = [0,\ 1]\,, \tag{24}$$

with the elements repeating $n_c$ and $n_e$ times, respectively. Then, $Y$, $\delta$, and $x$ form the preliminary simulated data $D_S$, serving as a starting point for optimization. For example, assume that we have the following: $n_c = 8$, $n_e = 9$, $v_c = 3$, $v_e = 4$, $t_{\max} = 20$. Then the data could look as shown in Table 1.

Subsequently, summary statistics for the simulated data are calculated. Possibilities are $\mathrm{KM}_c$, $\mathrm{CI}\left(\mathrm{KM}_c\right)_{LB}$, $\mathrm{CI}\left(\mathrm{KM}_c\right)_{UB}$, $\mathrm{KM}_e$, $\mathrm{CI}\left(\mathrm{KM}_e\right)_{LB}$, $\mathrm{CI}\left(\mathrm{KM}_e\right)_{UB}$, HR, $\mathrm{CI}\left(\mathrm{HR}\right)_{LB}$, and $\mathrm{CI}\left(\mathrm{HR}\right)_{UB}$. However, only a subset of the nine possibilities must be calculated, namely those that are also reported in the article and that will therefore be used for data simulation.

The subsequent optimization procedure involves $n_c + n_e$ parameters, which we collectively call $\xi$. Thus, each case $i$ in $D_S$ is coupled with one parameter $\xi_i$ that must be estimated. At iteration $q$, $Y_i$ is calculated as:

$$Y_i^q = e^{\xi_i} Y_i^{q-1}. \tag{25}$$

Here, $\xi_i$ is restricted to range between $\log\left[1/Y_i^{q-1}\right]$ and $\log\left[t_{\max}/Y_i^{q-1}\right]$; this ensures that the newly calculated observed response $Y_i^q$ is not lower than 1 and not higher than $t_{\max}$. In essence, the optimization procedure attempts to adjust $Y$ in a way such that interim summary statistics match the actual summary statistics. Let $E$ be a vector of all known summary statistics and $O$ be a vector (in fact, a function of $\xi$ and $D_S$) with the same kinds of summary statistics as $E$ but calculated from the simulated data $D_S$. To estimate $\xi$, we iteratively minimize the following loss function:

$$\phi\left(\xi,\ D_S\right) = \log\left[\sum_{r=1}^{|E|}\left[\left(\frac{O_r - E_r}{E_r} w_r\right)^2\right]\frac{1}{|E|}\right], \tag{26}$$

where $|E|$ is the number of used summary statistics. $w$ is a weight vector that we address below. In essence, we define the loss function as the log of the mean squared deviations

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



**Table 1**

*Mock data.*

| Subject | $Y$ | $\delta$ | $x$ |
|---|---|---|---|
| 1 | 7 | 0 | 0 |
| 2 | 17 | 0 | 0 |
| 3 | 10 | 0 | 0 |
| 4 | 20 | 0 | 0 |
| 5 | 6 | 0 | 0 |
| 6 | 20 | 1 | 0 |
| 7 | 3 | 1 | 0 |
| 8 | 3 | 1 | 0 |
| 9 | 11 | 0 | 1 |
| 10 | 10 | 0 | 1 |
| 11 | 6 | 0 | 1 |
| 12 | 2 | 0 | 1 |
| 13 | 6 | 0 | 1 |
| 14 | 8 | 1 | 1 |
| 15 | 4 | 1 | 1 |
| 16 | 8 | 1 | 1 |
| 17 | 18 | 1 | 1 |

between the observed and the expected summary statistics, scaled by the expected summary statistics (akin to the classical $\chi^2$ test statistic) and weighted by $w$. The scaling is done because the different kinds of summary statistics are on different scales and the weighting is done because different kinds of summary statistics might contribute more or less strongly to the accuracy of the resulting Bayes factors.

Remembering that $\phi(\xi,\ D_S)$ is a $n_c + n_e$-variate function and that calculating $O$

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



involves complex formulas, it becomes clear that the loss function in Equation 26 is very difficult to differentiate. Therefore, gradient-based optimization techniques cannot be used; instead, we rely on a derivative-free optimization tool called *Particle Swarm Optimization* (PSO; Kennedy & Eberhart, 1995; Shi & Eberhart, 1998) to minimize Equation 26. A detailed treatment of PSO is beyond the scope of this article; we refer the interested reader to Clerc (2006). We implemented PSO in R (R Core Team, 2022) using the "psoptim()" function of the "pso" R package (Bendtsen, 2022), keeping almost all default settings. Exceptions are the maximum number of PSO iterations and the allowed number of PSO iterations, which do not result in a decrease in the loss. The choice of our defaults is based on simulations, described below, that had the purpose of tuning these PSO parameters. Nevertheless, all arguments in the "psoptim()" function can be set as desired by the user.

### Tuning of PSO Parameters

For the tuning of some PSO parameters, we made use of three example data sets: the Kidney (McGilchrist & Aisbett, 1991), Lung (Loprinzi et al., 1994), and Colon (Laurie et al., 1989) data sets that are available through the "survival" R package (Therneau, 2021). The Kidney data set provides times until infection after insertion of a catheter in kidney patients. The Lung data set describes survival times of patients with advanced lung cancer. Lastly, the Colon data set presents recurrence and death times in patients receiving adjuvant chemotherapy for colon cancer. Here, we chose to only examine death as an endpoint. For all three example data sets we used sex as the independent variable, with males being coded as 0 and females as 1. The Kidney, Lung, and Colon data sets have sample sizes of 76, 228, and 929, respectively. For the KM median survival times and HR we calculated the corresponding 95% confidence intervals.

Importantly, we only used these data sets for the purpose of tuning the weights of summary statistics and the number of iterations in PSO. Therefore, no inferences from our results should be drawn.

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



### *Weights of Summary Statistics*

The weight vector $w$ in Equation 26 determines how influential certain summary statistics are in the calculation of the loss function, as well as for simulating data that yield a Bayes factor that is as close as possible to the true Bayes factor of the actual data set. We simulated 100 data sets for each of eight different sets of weights for the three example data sets. A Bayes factor was calculated for each simulated data set. We used a Normal prior with a mean of $\mu = 0$ and a standard deviation of $\sigma = 1$ for the $\beta$ parameter. These simulated Bayes factors were then compared to the Bayes factor that we calculated based on the full data set, using the same computational procedure as illustrated in the previous section.

Figures 2, 3, and 4 show the results, with each panel representing a different set of weights. For all three example data sets, a weight set in which the KM median survival times and the corresponding confidence intervals are not considered and HR is weighted twice as much as the corresponding HR confidence interval boundaries yields Bayes factors with the smallest variance. Moreover, there is almost no bias in the distribution of Bayes factors for the Lung and Colon data sets. A small amount of bias was found for the Kidney data set, which had the smallest sample size (see Figure 2). Using KM measures as well increases the variance and bias of the resulting Bayes factors. Consequently, it seems reasonable to ignore the KM estimates and instead only use HR and the corresponding confidence interval. In case the HR confidence interval is not mentioned in the original article, the results in Figures 2, 3, and 4 suggest that only using HR yields Bayes factors that are reasonable approximations to the true Bayes factor. Due to these results, we henceforth only consider HR and the corresponding confidence interval as potential summary statistics. Further, using only HR and the corresponding confidence interval has the additional advantage that the maximum possible response time $t_{max}$ becomes irrelevant when simulating data.

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



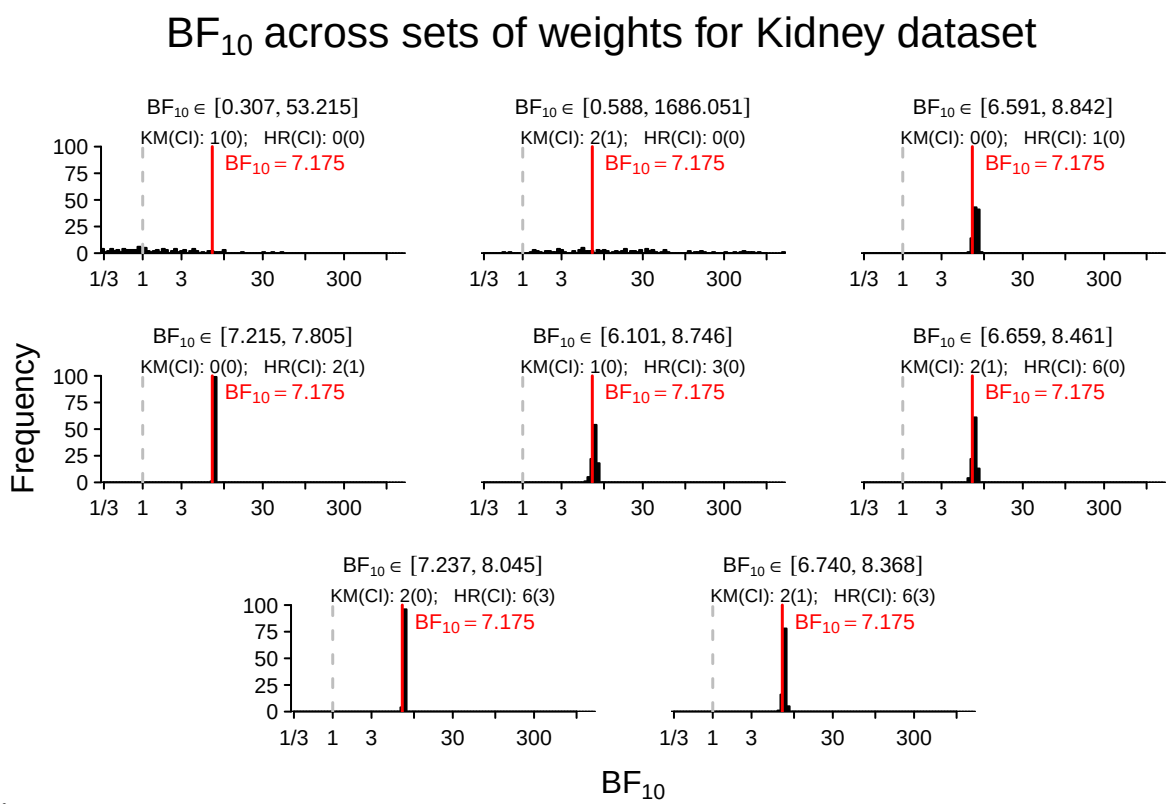


**Figure 2**

*Distribution of $BF_{10}$ for the Kidney data set. Panels display $BF_{10}$ for* 100 *simulated data sets using different sets of weights for summary statistics. The specific weights are printed in each panel, where KM represents $KM_c$ and $KM_e$, CI represents $CI(KM_c)_{LB}$, $CI(KM_c)_{UB}$, $CI(KM_e)_{LB}$, and $CI(KM_e)_{UB}$, HR represents HR, and CI represents $CI(HR)_{LB}$ and $CI(HR)_{UB}$. The red vertical line represents $BF_{10}$ for the full Kidney data.*

### *Maximum Number of PSO Iterations*

Another parameter of interest is the required number of PSO iterations for a satisfactory loss, bias, and variance of Bayes factors. This is especially important because the PSO algorithm is quite time-consuming. The higher the sample size, the larger the number of parameters in PSO and the longer the running time of PSO. To determine an approximate minimum number of PSO iterations, we simulated 100 data sets for each of six different maximum numbers of PSO iterations (i.e., 10, 30, 100, 300, 1,000, and 3,000)

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



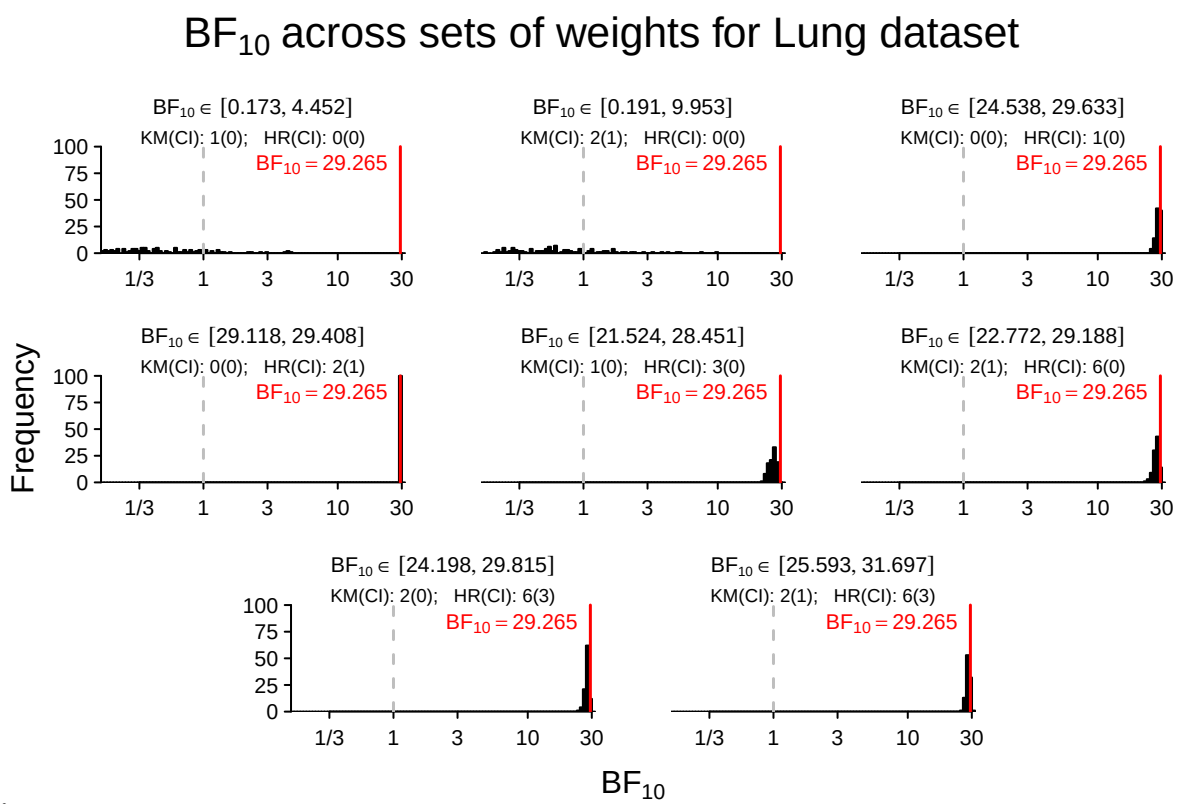


**Figure 3**

*Distribution of $BF_{10}$ for the Lung data set. Panels display $BF_{10}$ for 100 simulated data sets using different sets of weights for summary statistics. The specific weights are printed in each panel, where KM represents $KM_c$ and $KM_e$, CI represents $CI(KM_c)_{LB}$, $CI(KM_c)_{UB}$, $CI(KM_e)_{LB}$, and $CI(KM_e)_{UB}$, HR represents HR, and CI represents $CI(HR)_{LB}$ and $CI(HR)_{UB}$. The red vertical line represents $BF_{10}$ for the full Lung data.*

and calculated the Bayes factor. This was repeated for each of the three example data sets. The PSO algorithm stops either when the maximum number of PSO iterations is reached or when no reduction in loss is obtained within one fifth of the maximum number of PSO iterations (e.g., no improvement in 200 iterations when the maximum number of PSO iterations is $1,000$). Here as well, we used a Normal prior with a mean of $\mu = 0$ and a standard deviation of $\sigma = 1$ for the $\beta$ parameter.

Figure 5 shows the results for the case where only HR is used as a summary statistic and Figure 6 shows the results for the case where HR and the corresponding confidence

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



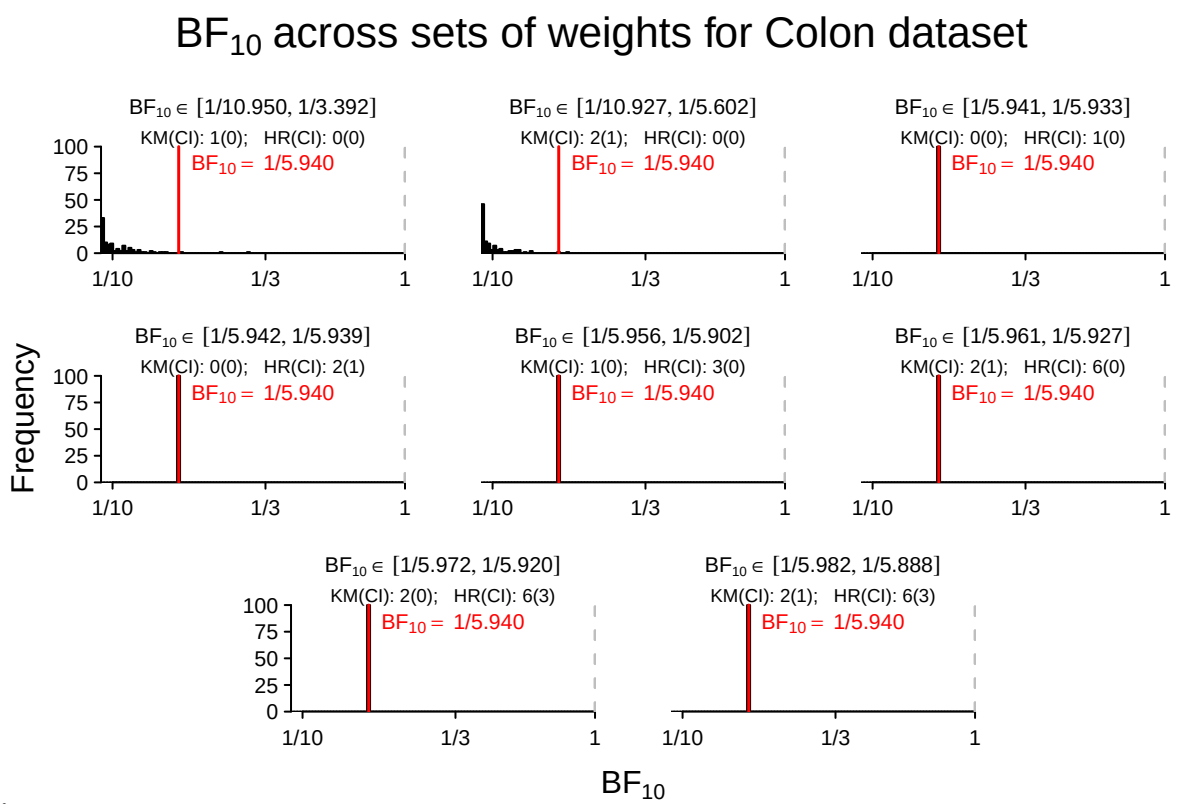


**Figure 4**

*Distribution of $BF_{10}$ for the Colon data set. Panels display $BF_{10}$ for* 100 *simulated data sets using different sets of weights for summary statistics. The specific weights are printed in each panel, where KM represents $KM_c$ and $KM_e$, CI represents $CI(KM_c)_{LB}$, $CI(KM_c)_{UB}$, $CI(KM_e)_{LB}$, and $CI(KM_e)_{UB}$, HR represents HR, and CI represents $CI(HR)_{LB}$ and $CI(HR)_{UB}$. The red vertical line represents $BF_{10}$ for the full Colon data.*

interval are used as summary statistics. Figure 5 indicates that only a small number of PSO iterations is required when only using HR because the variance of the Bayes factors does not improve significantly when more than approximately 30 or 100 iterations are used. In contrast, Figure 6 suggests that when using both HR and the corresponding confidence interval the variance of the Bayes factors decreases the more PSO iterations are used. To obtain a reasonable trade-off between the running time of PSO and the Bayes factor variance, we recommend running between 100 and 300 iterations.

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



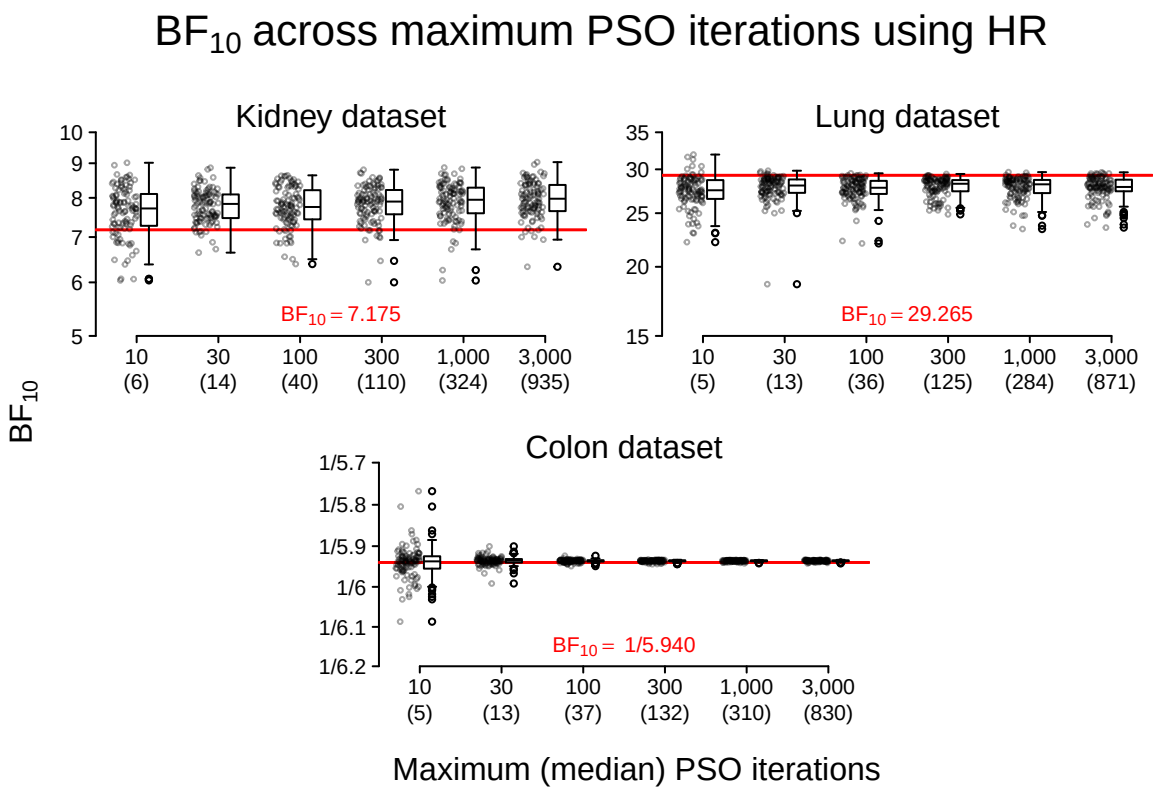


**Figure 5**

*Distribution of $BF_{10}$ applied to* 100 *simulated data sets using different maximum numbers of PSO iterations for the Kidney, Lung, and Colon data sets. HR is used for data simulation. Note that even though a maximum allowed number of PSO iterations was defined, optimization could stop earlier in case no improvement was found within one fifth of the maximum allowed number of iterations. Therefore, the median actual number of PSO iterations is given in parentheses on the x-axis. The red horizontal line represents $BF_{10}$ for the actual data set.*

## Reanalysis of Beigel et al. (2020)

To demonstrate our proposed procedure for simulating data from summary statistics and calculating one Bayes factor for each simulated data set, we will again reanalyze the study reported in Beigel et al. (2020). But this time we will only use the summary statistics reported in their article. As in the previous section, our reanalysis of

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



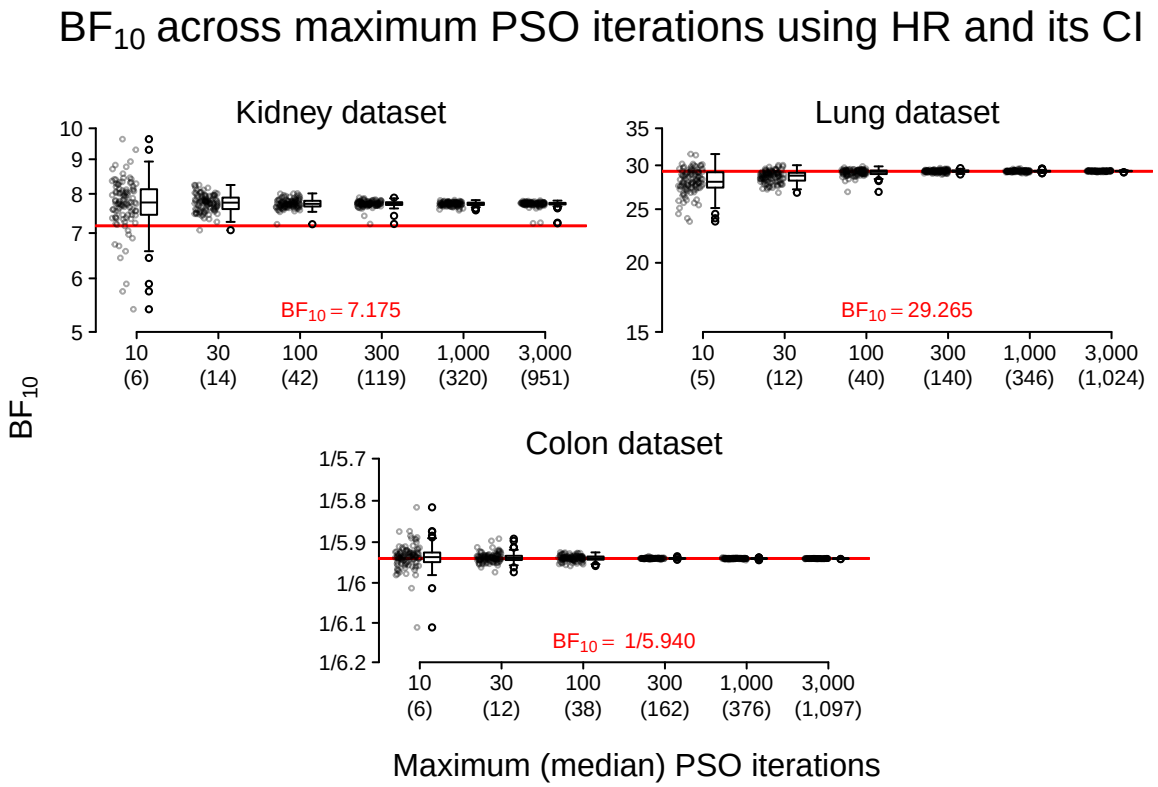


**Figure 6**

*Distribution of $BF_{10}$ applied to* 100 *simulated data sets using different maximum numbers of PSO iterations for the Kidney, Lung, and Colon data sets. HR and its* 95%*CI are used for data simulation. Note that even though a maximum allowed number of PSO iterations was defined, optimization could stop earlier in case no improvement was found within one fifth of the maximum allowed number of iterations. Therefore, the median actual number of PSO iterations is given in parentheses on the x-axis. The red horizontal line represents $BF_{10}$ for the actual data set.*

Beigel et al. (2020) omitted the stratification by actual disease severity and we used Efron's instead of Breslow's approximation to the true partial likelihood. Due to these small deviations, our results (i.e., HR = 1.312, 95% CI = [1.136, 1.514]) are slightly different from the results reported in Beigel et al. (2020) (i.e., HR = 1.29, 95% CI = [1.12, 1.49]). In the following, we use our calculated summary statistics as if they were the only results provided in Beigel et al. (2020).

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



Using the "baymedr" R package, we simulated 100 data sets based on the summary statistics as follows:

```
sim_data <- coxph_data_sim(            # See ?coxph_data_sim for details
    n_data = 100,                      # Number of data sets to be simulated
    ns_c   = 521,                      # Sample size (control condition)
    ns_e   = 541,                      # Sample size (experimental condition)
    ne_c   = 352,                      # Number of events (control condition)
    ne_e   = 399,                      # Number of events (experimental
                                       #     condition)
    cox_hr = c(1.312, 1.136, 1.514),   # HR, lower bound CI, upper bound CI
    cox_hr_ci_level = 0.95,            # Confidence level CI
    maxit  = 300,                      # Max number of PSO iterations (for
                                       #     psoptim())
    maxit.stagnate = ceiling(300 / 5), # Max number of PSO iterations without
                                       #     reduction in loss (for psoptim())
    cores  = 5                         # Number of cores to be used
)
```

Execution of this code creates object `sim_data`, which is a list with one entry for each simulated data set. Each entry contains the simulated full data set and additional output related to PSO (e.g., the $\xi$ parameter values, the achieved loss, and the number of iterations).

Subsequently, we calculated one Bayes factor for each of the 100 simulated data sets. We used a truncated (because we have a one-sided alternative hypothesis) Normal prior for $\beta$ with a mean of $\mu = 0$ and a standard deviation of $\sigma = 1$:

```
sim_bf <- coxph_bf(            # See ?coxph_bf for details
    data        = sim_data,    # Object containing the data
```

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



```
    null_value  = 0,           # H0 value
    alternative = "one.sided", # H1 type (one- or two-sided)
    direction   = "high",      # H1 direction (low or high)
    prior_mean  = 0,           # Beta prior mean
    prior_sd    = 1            # Beta prior SD
)
```

The result is object `sim_bf` that contains all relevant information of the Bayes factor analysis. When typing the name of the object (i.e., `sim_bf`) into the R console, a concise summary is provided, which is almost equivalent to the one shown in the previous section. Since we now have multiple Bayes factors instead of only one, the output summarizes the resulting Bayes factors with a median and the median absolute deviation that is on a similar scale as the traditional standard deviation (Gelman et al., 2020; Hampel, 1974; Huber, 1981):

```
******************************
Cox proportional hazards analysis
---------------------------------
H0:               beta == 0
H+:               beta > 0
Normal prior:     Mean = 0.000
                  SD = 1.000

    Median BF+0 = 136.601
    MAD SD BF+0 = 0.370
******************************
```

A histogram of the resulting Bayes factors can be found in Figure 7. The simulated Bayes factors range between $\mathrm{BF}_{+0} = 133.6$ and $\mathrm{BF}_{+0} = 135.7$, thus supporting the

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



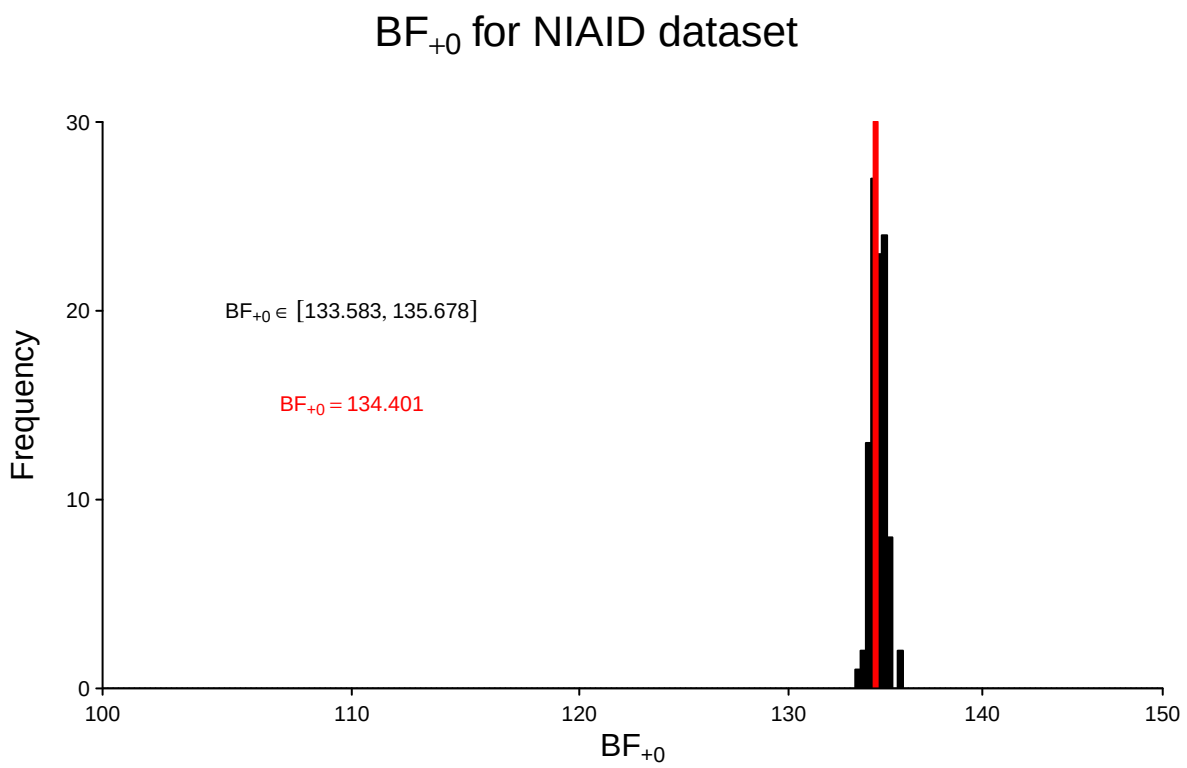


**Figure 7**

*Distribution of $BF_{+0}$ applied to* 100 *simulated data sets for the data set described in Beigel et al. (2020). HR and its* 95%*CI are used for data simulation. The red vertical line represents $BF_{+0}$ for the full data. See text for details.*

conclusion of Beigel et al. (2020) that Remdesivir seems to have a beneficial effect on the recovery of patients with Covid-19. The red vertical line represents the actual Bayes factor that was obtained in the previous section where we used the full data of Beigel et al. (2020). As such, we can conclude that our approximate Bayes factor based on summary statistics is virtually unbiased (i.e., the red line is in the middle of the black histogram) and has very low variability (i.e., the histogram occupies a very limited range on the x-axis).

In the next section, we compare our approach for calculating Bayes factors to an approximation approach advocated by Bartoš and Wagenmakers (2022).

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



## Comparison with Savage-Dickey Normal Approximation

In the previous two sections, we demonstrated how Bayes factors for Cox models can be calculated based on the full data and based on summary statistics. For a real data set, we showed that the distribution of Bayes factors resulting from simulated data based on summary statistics approximate the Bayes factor that is calculated based on the full data very well. In the following, we compare our procedure for calculating Bayes factors with an alternative procedure that tries to approximate Bayes factors.

Bartoš and Wagenmakers (2022) introduced a generic method that uses a Normal approximation of the likelihood function to calculate a Bayes factor for various statistical designs. If $\mathcal{H}_0$ is a point null hypothesis, the Bayes factor is the ratio of the ordinate of the prior (i.e., density or height of the prior) and the ordinate of the posterior for the parameter of interest $\beta$, evaluated at the null value $\beta_0$. This ratio is called the Savage-Dickey density ratio (e.g., Dickey & Lientz, 1970):

$$\mathrm{BF}_{10} = \frac{f\left(\beta = \beta_0 \mid \mathcal{H}_1\right)}{f\left(\beta = \beta_0 \mid D,\ \mathcal{H}_1\right)}. \tag{27}$$

For this to work, only the maximum likelihood estimate and the corresponding standard error of the underlying likelihood function of the respective statistical analysis must be known ($\hat{\beta}$ and $SE\left(\hat{\beta}\right)$, respectively). If the prior for the parameter of interest is defined as a Normal distribution with mean $\mu$ and variance $\sigma^2$, a closed-form solution for the Bayes factor is available (cf. the last equation on p. 3 of Bartoš & Wagenmakers, 2022):

$$\mathrm{BF}_{01} = \sqrt{\frac{\sigma^2 + SE\left(\hat{\beta}\right)^2}{SE\left(\hat{\beta}\right)^2}} \exp\left(-\frac{1}{2}\left[\frac{\left(\hat{\beta} - \beta_0\right)^2}{SE\left(\hat{\beta}\right)^2} - \frac{\left(\hat{\beta} - \mu\right)^2}{\sigma^2 + SE\left(\hat{\beta}\right)^2}\right]\right). \tag{28}$$

Bartoš and Wagenmakers (2022) use the examples of a two-sample $t$-test, a parametric survival analysis, and a meta-regression to demonstrate that their approximate Bayes factors are accurate and can be applied to a wide range of statistical models.

We investigated whether the method by Bartoš and Wagenmakers (2022) also yields accurate Bayes factors for semi-parametric Cox models and how they compare to the Bayes

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



factors resulting from our method. This is important because the method by Bartoš and Wagenmakers (2022) provides a closed-form solution for calculating Bayes factors from $\hat{\beta}$ and $SE\left(\hat{\beta}\right)$ directly. In other words, there is no need for simulating data from summary statistics, which makes their method time-efficient. As such, if both methods were to provide equally accurate Bayes factors, the method by Bartoš and Wagenmakers (2022) would be preferable. We took the three example data sets mentioned before and calculated one Bayes factor through our method and one Bayes factor through the method by Bartoš and Wagenmakers (2022), to mimic the situation where full data sets are available. Moreover, we simulated 100 data sets for each example data set using the corresponding summary statistics of the three data sets and calculated one Bayes factor for each simulated data set using our method.

The results are shown in Figure 8, where the red vertical line represents the Bayes factor for the full data set, which is calculated using our method, the blue vertical line represents the Bayes factor resulting from the Savage-Dickey Normal approximation method advocated in Bartoš and Wagenmakers (2022), and the histogram shows Bayes factors from our method when using summary statistics. The Bayes factors resulting from the method by Bartoš and Wagenmakers (2022) are qualitatively similar to the true Bayes factors and can be used when a rough approximation is acceptable. However, when precise estimates are desirable, our method is preferable, both in the scenario where the full data set is available and in the scenario where only summary statistics are available. For the user it is a trade-off between accuracy and computation time. As shown in Figure 8, our Bayes factors are much more accurate. However, in the case where summary statistics must be used, our approach has a computation time that is orders of magnitude higher than the approach by Bartoš and Wagenmakers (2022).

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



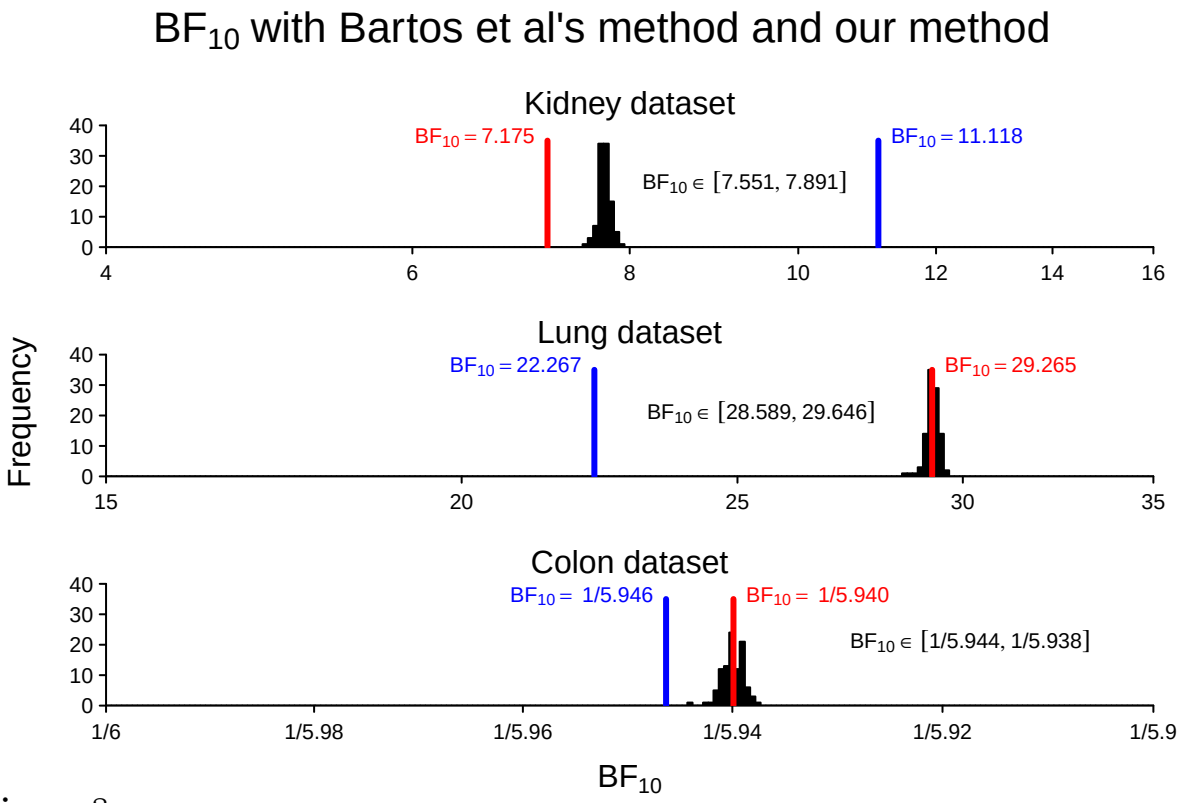


**Figure 8**

*Distribution of $BF_{10}$ applied to* 100 *simulated data sets for the Kidney, Lung, and Colon data sets using our approach. HR and its* 95%*CI are used for data simulation. The red vertical line represents $BF_{10}$ for the full data set using our approach. The blue vertical line represents $BF_{10}$ for the approximation by Bartoš and Wagenmakers (2022).*

## Discussion

The analysis of time-to-event data is commonly applied in biomedical research and provides important insights into the effectiveness of therapies. Most often, Cox regression (Cox, 1972) is used to analyze these kinds of data and NHST is then applied in order to make inferences. As an alternative to NHST, we presented a procedure to calculate Bayes factors for simple Cox models and offered the R package "baymedr" (Linde et al., 2022) as an easy-to-use implementation. "baymedr" can be used to calculate a Bayes factor for full data and to simulate multiple Bayes factors based on summary statistics as reported in

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



articles.

Our procedure for calculating Bayes factors for Cox models is oriented towards analysis strategies that seem prevalent in the biomedical literature: the semi-parametric Cox regression comparing a treatment to a control (or different treatment) condition. The use of Bayes factors specifically allows the important contrast between evidence that an effect is present and evidence than an effect is absent.

At the same time, many features and functionalities are still missing. For example, it would be desirable to also make the Cox partial likelihood (Cox, 1972) and Breslow's approximation to the true partial likelihood (Breslow, 1974) available. Moreover, the procedure implemented in "baymedr" should allow researchers to calculate Bayes factors for more complex Cox models. This includes, for instance, allowing for more than one independent variable, whether it be discrete or continuous, and allowing for stratification. Such an extension to more than one independent variable is not straightforward, as it will not be possible to calculate the Bayes factor through Gaussian quadrature. Instead, one of many more time-consuming approaches would have to be employed. For instance, the posterior distribution could be estimated through MCMC sampling (e.g., Betancourt, 2018; Brooks et al., 2011; Gilks et al., 1995; van Ravenzwaaij et al., 2018); the posterior samples could then be used to estimate the marginal likelihood through bridge sampling (e.g., Gronau et al., 2017).

## Conclusion

Cox proportional hazards regression is commonly used to analyze time-to-event data in biomedical research. Typically, the frequentist framework is used to make inferences. We provided a procedure for calculating Bayes factors for simple Cox models that can be applied both to the full data set and to summary statistics. The latter could be considered especially important because it allows reanalyzing multiple existing studies to make judgments and decisions about the effectiveness of therapies. We offered "baymedr" (Linde

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



et al., 2022), an R package that is aimed at all researchers desiring to calculate a Bayes factor for their Cox regression.

## Acknowledgments

This research was supported by a Dutch scientific organization VIDI fellowship grant (016.Vidi.188.001) to Don van Ravenzwaaij and a Japanese JSPS KAKENHI grant awarded to Jorge N. Tendeiro (21K20211). We thank Merle-Marie Pittelkow for consulting us on which summary statistics are reported most often in articles that use a Cox proportional hazards regression.

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



## References

Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A (General)*, *132*(2), 235–244.

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

Bartoš, F. (2022). RoBSA: An R package for robust Bayesian survival analyses [R package version 1.0.0]. https://CRAN.R-project.org/package=RoBSA

Bartoš, F., & Wagenmakers, E.-J. (2022). Fast and accurate approximation to informed Bayes factors for focal parameters.

Beigel, J. H., Tomashek, K. M., Dodd, L. E., Mehta, A. K., Zingman, B. S., Kalil, A. C., Hohmann, E., Chu, H. Y., Luetkemeyer, A., Kline, S., Lopez de Castilla, D., Finberg, R. W., Dierberg, K., Tapson, V., Hsieh, L., Patterson, T. F., Paredes, R., Sweeney, D. A., Short, W. R., … Lane, C. (2020). Remdesivir for the treatment of Covid-19 - final report. *The New England Journal of Medicine*, *383*(19), 1813–1826. https://doi.org/10.1056/NEJMoa2007764

Bendtsen, C. (2022). *pso: Particle swarm optimization* [R package version 1.0.4]. https://CRAN.R-project.org/package=pso

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, *2*(3), 317–335. https://doi.org/10.1214/ss/1177013238

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. https://doi.org/10.2307/2289131

Betancourt, M. (2018). A conceptual introduction to Hamiltonian Monte Carlo.

Bradburn, M. J., Clark, T. G., Love, S. B., & Altman, D. G. (2003a). Survival analysis part II: Multivariate data analysis - an introduction to concepts and methods. *British Journal of Cancer*, *89*, 431–436. https://doi.org/10.1038/sj.bjc.6601119

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



Bradburn, M. J., Clark, T. G., Love, S. B., & Altman, D. G. (2003b). Survival analysis part III: Multivariate data analysis - choosing a model and assessing its adequacy and fit. *British Journal of Cancer*, *89*, 605–611. https://doi.org/10.1038/sj.bjc.6601120

Brent, R. P. (1973). *Algorithms for minimization without derivatives*. Prentice Hall.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, *30*(1), 89–99. https://doi.org/10.2307/2529620

Brooks, S., Gelman, A., Jones, G. L., & Meng, X.-L. (2011). *Handbook of Markov chain Monte Carlo*. Chapman & Hall/CRC.

Chalmers, I., & Glasziou, P. (2009). Avoidable waste in the production and reporting of research evidence. *The Lancet*, *374*(9683), 86–89. https://doi.org/10.1016/S0140-6736(09)60329-9

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. https://doi.org/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. https://doi.org/10.1016/j.jhep.2007.02.015

Clark, T. G., Bradburn, M. J., Love, S. B., & Altman, D. G. (2003a). Survival analysis part I: Basic concepts and first analyses. *British Journal of Cancer*, *89*, 232–238. https://doi.org/10.1038/sj.bjc.6601118

Clark, T. G., Bradburn, M. J., Love, S. B., & Altman, D. G. (2003b). Survival analysis part IV: Further concepts and methods in survival analysis. *British Journal of Cancer*, *89*, 781–786. https://doi.org/10.1038/sj.bjc.6601117

Clerc, M. (2006). *Particle swarm optimization*. ISTE.

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

Collett, D. (2015). *Modelling survival data in medical research* (3rd). CRC Press.

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



Cook, J. D. (2012). *Avoiding underflow in Bayesian computations*. Retrieved August 23, 2022, from https://www.johndcook.com/blog/2012/07/26/avoiding-underflow-in-bayesian-computations/

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

Dean, N. C., Silver, M. P., Bateman, K. A., James, B., Hadlock, C. J., & Hale, D. (2001). Decreased mortality after implementation of a treatment guideline for community-acquired pneumonia. *The American Journal of Medicine*, *110*(6), 451–457. https://doi.org/10.1016/S0002-9343(00)00744-0

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. https://doi.org/10.1214/aoms/1177697203

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

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, *72*(359), 557–565.

Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology*, *63*, 665–694. https://doi.org/10.1348/000711010X502733

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

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

Gelman, A., Hill, J., & Vehtari, A. (2020). *Regression and other stories*. Cambridge University Press.

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606. https://doi.org/10.1016/j.socec.2004.09.033

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

Goldberg, D. (1991). What every computer scientist should know about floating-point arithmetic. *ACM Computing Surveys*, *23*(1), 5–48. https://doi.org/10.1145/103162.103163

Goodman, S. N. (1999a). Toward evidence-based medical statistics. 1: The p value fallacy. *Annals of Internal Medicine*, *130*(12), 995–1004. https://doi.org/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. https://doi.org/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. https://doi.org/10.1053/j.seminhematol.2008.04.003

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., & Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, *81*, 80–97. https://doi.org/10.1016/j.jmp.2017.09.005

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, *69*, 383–393. https://doi.org/10.1080/01621459.1974.10482962

Harrell, F. E. (2015). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis* (2nd). Springer.

Harrell, F. E. (2022). *rms: Regression modeling strategies* [R package version 6.3-0]. https://CRAN.R-project.org/package=rms

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



Hosmer, D. W., Lemeshow, S., & May, S. (2008). *Applied survival analysis: Regression modeling of time-to-event data* (2nd). John Wiley & Sons.

Huber, P. J. (1981). *Robust statistics*. John Wiley & Sons, Inc.

JASP Team. (2022). JASP (Version 0.16.3)[Computer software]. https://jasp-stats.org/

Jeffreys, H. (1939). *Theory of probability*. The Clarendon Press.

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

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

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53*(282), 457–481. https://doi.org/10.1080/01621459.1958.10501452

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

Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. *Proceedings of ICNN'95 - International Conference on Neural Networks*, *4*, 1942–1948. https://doi.org/10.1109/ICNN.1995.488968

Keysers, C., Gazzola, V., & Wagenmakers, E.-J. (2020). Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nature Neuroscience*, *23*, 788–799. https://doi.org/10.1038/s41593-020-0660-4

Klein, J. P., & Moeschberger, M. L. (1997). *Survival analysis: Techniques for censored and truncated data*. Springer.

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

Laurie, J. A., Moertel, C. G., Fleming, T. R., Wieand, H. S., Leigh, J. E., Rubin, J., McCormack, G. W., Gerstner, J. B., Krook, J. E., & Malliard, J. (1989). Surgical adjuvant therapy of large-bowel carcinoma: An evaluation of levamisole and the combination of levamisole and fluorouracil. The North Central Cancer Treatment

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



Group and the Mayo Clinic. *Journal of Clinical Oncology*, *7*(10), 1447–1456. https://doi.org/10.1200/JCO.1989.7.10.1447

Leung, K.-M., Elashoff, R. M., & Afifi, A. A. (1997). Censoring issues in survival analysis. *Annual Review of Public Health*, *18*, 83–104. https://doi.org/10.1146/annurev.publhealth.18.1.83

Linde, M., van Ravenzwaaij, D., & Tendeiro, J. N. (2022). *baymedr: Computation of Bayes factors for common biomedical designs* [R package version 0.1.1.9000]. https://github.com/maxlinde/baymedr

Loprinzi, C. L., Laurie, J. A., Wieand, H. S., Krook, J. E., Novotny, P. J., Kugler, J. W., Bartel, J., Law, M., Bateman, M., & Klatt, N. E. (1994). Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. *Journal of Clinical Oncology*, *12*(3), 601–607. https://doi.org/10.1200/JCO.1994.12.3.601

McGilchrist, C. A., & Aisbett, C. W. (1991). Regression with frailty in survival analysis. *Biometrics*, *47*(2), 461–466. https://doi.org/10.2307/2532138

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, *73*(sup1), 235–245. https://doi.org/10.1080/00031305.2018.1527253

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes factors for common designs* [R package version 0.9.12-4.2]. https://CRAN.R-project.org/package=BayesFactor

Pearse, R. M., Moreno, R. P., Bauer, P., Pelosi, P., Metnitz, P., Spies, C., Vallet, B., Vincent, J.-L., Hoeft, A., & Rhodes, A. (2012). Mortality after surgery in Europe: A 7 day cohort study. *The Lancet*, *380*(9847), 1059–1065. https://doi.org/10.1016/S0140-6736(12)61148-9

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. https://doi.org/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. https://doi.org/10.3758/PBR.16.2.225

Sanborn, A. N., & Hills, T. T. (2014). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic Bulletin & Review*, *21*(2), 283–300. https://doi.org/10.3758/s13423-013-0518-9

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin and Review*, *25*(1), 128–142. https://doi.org/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. https://doi.org/10.1037/met0000061

Senn, S. (2008). *Statistical issues in drug development* (2nd). John Wiley & Sons.

Shi, Y., & Eberhart, R. (1998). A modified particle swarm optimizer. *1998 IEEE International Conference on Evolutionary Computation Proceedings. IEEE World Congress on Computational Intelligence*, 69–73. https://doi.org/10.1109/ICEC.1998.699146

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

Sutton, A. J., & Abrams, R., Keith. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research*, *10*(4), 277–303. https://doi.org/10.1177/096228020101000404

Tendeiro, J. N., & Kiers, H. A. L. (2019). A review of issues about null hypothesis Bayesian testing. *Psychological Methods*, *24*(6), 774–795. https://doi.org/10.1037/met0000221

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



Tendeiro, J. N., Kiers, H. A. L., & Van Ravenzwaaij, D. (2022). Worked-out examples of the adequacy of Bayesian optional stopping. *Psychonomic Bulletin & Review*, *29*(1), 70–87. https://doi.org/10.3758/s13423-021-01962-5

Therneau, T. M. (2021). *A package for survival analysis in R* [R package version 3.2-13]. https://CRAN.R-project.org/package=survival

Therneau, T. M., & Grambsch, P. M. (2000). *Modeling survival data: Extending the Cox model*. Springer.

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. https://doi.org/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. https://doi.org/10.1371/journal.pone.0173184

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. https://doi.org/10.1186/s12874-019-0699-7

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

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

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. part I: Theoretical advantages and

medRxiv preprint doi: https://doi.org/10.1101/2022.11.02.22281762; this version posted November 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3

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. https://doi.org/10.1080/00031305.2016.1154108

Wickham, H., Hester, J., & Chang, W. (2019). *devtools: Tools to make developing R packages easier* [R package version 2.2.0]. https://CRAN.R-project.org/package=devtools