Bayes Factors for Two-group Comparisons in Cox Regression with an Application for Reverse-Engineering Raw Data from Summary Statistics

Maximilian Linde[1,2], Jorge N. Tendeiro[3], and Don van Ravenzwaaij[2]

[1]GESIS - Leibniz Institute for the Social Sciences, Cologne, Germany

[2]Faculty of Behavioural and Social Sciences, University of Groningen, Groningen, The Netherlands

[3]Graduate School of Advanced Science and Engineering, Hiroshima University, Hiroshima, Japan

Author Note

Correspondence concerning this article should be addressed to: Maximilian Linde, GESIS - Leibniz Institute for the Social Sciences, Cologne, Germany, E-mail: maximilian.linde@gesis.org.

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

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 compute Bayes factors for simple Cox models, both for the scenario where the full data are 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, simulation, summary statistics, survival

Bayes Factors for Two-group Comparisons in Cox Regression with an Application for Reverse-Engineering Raw Data from Summary Statistics

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., 17, 29, 64]. 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., 24, 57]. 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 [37]. 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 11, 12, 18, 19, 22, 37, 39, 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., 60] 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 [2, 47, 73]. 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 [1, 59, 68].

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; 23]. This allows directly contrasting the evidence for the null hypothesis $\mathcal{H}_0$ that there is no effect with an alternative hypothesis $\mathcal{H}_1$ that operationalizes that there is some effect; it also allows

monitoring results and continuing or stopping data collection at will. Moreover, to the best of our knowledge, so far Cox regression can only be conducted when the full data are 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., 28, 67]. We propose a novel procedure that describes 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; 52] 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 compute a Bayes factor when the full data are 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 compute 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 [4].

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 (e.g., due to the study ending before the event is observed, participants withdrawing from the study, or failure to follow-up with participants), leading to right-censored observations [e.g., 37, 39, 48, 51]. 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 probability that a participant is still alive

after a particular time t is given by the *survival function*:

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

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 have survived until time t [cf. 18]:

$$\lambda(t) = \lim_{u \to 0} \frac{P(t < T \leq t + u \mid T > t)}{u}. \tag{2}$$

Various kinds of survival analyses exist for estimating $S(t)$ and $\lambda(t)$. The semi-parametric Cox regression [23] 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 $i \in \{1, \ldots, n\}$ where n represents the sample size, consists of the observed response Y_i and an event indicator δ_i, designating whether the event of interest occurred (1) or not (0).

We assume that we have one independent variable x that is dichotomous, indicating membership to one of two conditions (control = 0; experimental = 1). We refer to the combination of Y, δ, and x as the data D for a generic participant. The Cox model [23] is expressed as:

$$\lambda(t \mid x) = \lambda(t)\, e^{x\beta}, \tag{3}$$

where β is the parameter we aim to estimate. Assuming that $x_i \in \{0, 1\}$, the hazard ratio is:

$$\text{HR} = \frac{\lambda(t \mid x = 1)}{\lambda(t \mid x = 0)} = \frac{\lambda(t)\, e^{\beta}}{\lambda(t)} = e^{\beta}. \tag{4}$$

The estimation of β is based most often on maximum likelihood estimation. Commonly, either the original Cox partial likelihood [23], Breslow's approximation to the true partial likelihood [13], or Efron's approximation to the true partial likelihood [27] are used [e.g., 70]. 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 [69, 70] and the "rms" R package [38].

Once the model is estimated, a confidence interval for β or HR can be computed. Alternatively, null hypothesis significance testing (NHST) in the form of a Wald test (for instance) is conducted. 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.

The Bayes Factor

The use of NHST in biomedical research is ubiquitous [16] even though it has been criticized repeatedly [e.g., 8, 9, 21, 26, 31, 33, 34, 35, 55, 72, 75, 76, 77]. 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 $\mathcal{H}_0$ [60, 73, 75, 76]. 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 [59, 61, 62, 63, 68]. This might have the implication that fewer resources are wasted because neither too many nor too few cases are sampled [15]. In addition, the results of NHBT are easy to interpret and are arguably more in line with researchers' questions compared to NHST.

The most common vehicle of NHBT is the *Bayes factor* [41, 42, 43, 45], which quantifies the relative probabilities of the data under $\mathcal{H}_0$ and $\mathcal{H}_1$. For example, a Bayes factor of $\mathrm{BF}_{10} = 1$ indicates a perfect balance between $\mathcal{H}_0$ and $\mathcal{H}_1$, given the choice of prior and model. 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. Even though the Bayes factor is a continuous measure of evidence, several schemes have been proposed to classify Bayes factors into categories that represent qualitatively different degrees of evidence. Table 1 shows Bayes factor thresholds suggested by [43]. Later, [50] offered alternative interpretations for the thresholds established by [43]. An alternative classification scheme was proposed by [45], with the following thresholds:

BF_{10}	Jeffreys (1961)[43]	Lee & Wagenmakers (2013)[50]
1		No evidence
$1-3$	Not worth more than a bare mention	Anecdotal evidence
$3-10$	Substantial evidence	Moderate evidence
$10-30$	Strong evidence	Strong evidence
$30-100$	Very strong evidence	Very strong evidence
>100	Decisive evidence	Extreme evidence

Table 1

Bayes factor evidence thresholds with interpretations proposed by [43] and adaptations thereof proposed by [50]. All interpretations are in favor of $\mathcal{H}_1$; the same interpretations in favor of $\mathcal{H}_0$ apply to the inverse of the Bayes factors.

$1-3$: not worth more than a bare mention; $3-20$: positive; $20-150$: strong; >150: very strong. Importantly, all of these schemes should merely be considered as rules of thumb. They should be considered with caution and properly adapted to the problem at hand.

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:

$$\underbrace{\frac{P\left(\mathcal{H}_1 \mid D\right)}{P\left(\mathcal{H}_0 \mid D\right)}}_{\text{Posterior odds}} = \underbrace{\frac{P\left(D \mid \mathcal{H}_1\right)}{P\left(D \mid \mathcal{H}_0\right)}}_{\text{Bayes factor, } \mathrm{BF}_{10}} \times \underbrace{\frac{P\left(\mathcal{H}_1\right)}{P\left(\mathcal{H}_0\right)}}_{\text{Prior odds}}. \tag{5}$$

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 5 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 5 is computed by integrating the product of the prior distribution for the parameter of interest under $\mathcal{H}_1$ and the likelihood function. In the case of Cox regression with one dichotomous independent variable, the parameter of interest is β. Consequently, the numerator of the Bayes factor in Equation 5 is:

$$P\left(D \mid \mathcal{H}_1\right) = \int_{\beta \in \Omega_1} f\left(D \mid \beta\right) f\left(\beta\right) \mathrm{d}\beta, \tag{6}$$

where Ω_1 is the range of β parameter values under $\mathcal{H}_1$. If a point $\mathcal{H}_0$ is used, the denominator of the Bayes factor in Equation 5 is simply the density of the likelihood evaluated at the null value β_0:

$$P\left(D \mid \mathcal{H}_0\right) = f\left(D \mid \beta = \beta_0\right). \tag{7}$$

Even though the Bayes factor is independent of the prior odds, it is sensitive to the choice of prior for β [e.g., 30, 45, 66, 74].

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

$$\mathrm{BF}_{10} = \frac{\int_{\beta \in \Omega_1} f\left(D \mid \beta\right) f\left(\beta\right) \mathrm{d}\beta}{f\left(D \mid \beta = \beta_0\right)}, \tag{8}$$

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

The Need for Methods for Full Data and Summary Statistics

The computation 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 computing Bayes factors for various research designs. Examples are the R packages "BayesFactor" [56] and "baymedr" [52], and point-and-click software like "JASP" [40]. Moreover, for Bayesian parametric survival analysis, there is the "RoBSA" R package [3]. However, to the best of our knowledge, no software implementation currently exists that allows computing Bayes factors for Cox models. In addition to a module for computing 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 [52] to compute Bayes factors for Cox models. In the first section, we focus on the situation where the full data are 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 https://osf.io/37ut2/).

Computing a Bayes Factor from the Full Data

We applied our approach for computing Bayes factors for Cox models to an empirical data set, as described in [5]. The goal of this double-blind, randomized, placebo-controlled trial 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 either being dismissed from the hospital or remaining in hospital solely for the purpose of infection control.

[5] 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 $\mathcal{H}_1$ is used (i.e., a one-sided $\mathcal{H}_1$ 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. 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" [5, 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 in the article by 5].

Our reanalysis of [5] omitted the stratification by actual disease severity. Furthermore, we used Efron's instead of Breslow's approximation. 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 [5]: 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 [78] 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 compute a Bayes factor for the full data in [5]. The data must

have the survival time, the event indicator, and the independent variable in that order as columns. We used a truncated (because we have a one-sided $\mathcal{H}_1$) Normal prior for β 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 obtained Bayes factor of $\mathrm{BF}_{+0} = 134.401$ suggests that the data are 134 times more likely to have been generated under $\mathcal{H}_1$ compared to $\mathcal{H}_0$, given the choice of prior and model. Thus, according to approximate Bayes factor thresholds proposed by [43], we found decisive evidence ('extreme' according to [50] and 'strong' according to [45]) for the hypothesis that patients receiving Remdesivir recover quicker than patients receiving a placebo.

Computing Bayes Factors from Summary Statistics

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

When the full data are not available, an attractive alternative is to simulate data within the constraints of summary statistics that are known. Following such an approach, it is imperative to demonstrate that the simulated data are sufficiently constrained by the available summary statistics. To the best of our knowledge, our proposed approach for

computing Bayes factors from summary statistics is unique and novel. Therefore, we deem our procedure an important contribution for verifying results of published studies. 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; 44] median survival times within each condition, with corresponding confidence intervals, KM_c, $\mathrm{CI}\left(\mathrm{KM}_c\right)_{LB}$, $\mathrm{CI}\left(\mathrm{KM}_c\right)_{UB}$, and 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 [23], 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:

$$Y \sim \mathrm{Uniform}\left(1,\ t_{max}\right). \tag{9}$$

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 δ:

$$\delta = [0,\ 1,\ 0,\ 1], \tag{10}$$

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:

$$x = [0,\ 1], \tag{11}$$

with the elements repeating n_c and n_e times, respectively. Then, Y, δ, and x form the preliminary simulated data D_S, serving as a starting point for optimization.

Subsequently, summary statistics for the simulated data are computed. Possibilities are KM_c, $\mathrm{CI}\left(\mathrm{KM}_c\right)_{LB}$, $\mathrm{CI}\left(\mathrm{KM}_c\right)_{UB}$, 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 computed, 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 ξ. Thus, each case i in D_S is coupled with one parameter ξ_i that must be estimated. At iteration q, Y_i is computed as:

$$Y_i^q = e^{\xi_i} Y_i^{q-1}. \tag{12}$$

Here, ξ_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 computed 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 ξ and D_S) with the same kinds of summary statistics as E but computed from the simulated data D_S. To estimate ξ, we iteratively minimize the following loss function:

$$\phi\left(\xi,\ D_S\right) = \log\left[\sum_{r=1}^{|E|}\left[\left(\frac{O_r - E_r}{E_r} w_r\right)^2\right]\frac{1}{|E|}\right], \tag{13}$$

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 between the observed and the expected summary statistics, scaled by the expected summary statistics (akin to the classical χ^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 $\phi\left(\xi,\ D_S\right)$ is a $(n_c + n_e)$-variate function and that computing O involves complex formulas, it becomes clear that the loss function in Equation 13 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; 46, 65] to minimize Equation 13. A detailed treatment of PSO is beyond the scope of this article; we refer the interested reader to [20]. We implemented PSO in R [58] using the "psoptim()" function of the "pso" R package [7], 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. 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 [54], Lung [53], and Colon [49] data sets that are available through the "survival" R package [69]. 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 computed 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.

Weights of Summary Statistics

The weight vector w in Equation 13 determines how influential certain summary statistics are in the computation of the loss function. We simulated 100 data sets for each of eight different sets of weights for the three example data sets. A Bayes factor was

computed 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 β parameter. These simulated Bayes factors were then compared to the Bayes factor that we computed based on the full data set, using the same computational procedure as illustrated in the previous section.

The results are shown in Appendix A. 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 A1). Using KM measures as well increases the variance and bias of the resulting Bayes factors. Consequently, it seems reasonable to ignore the KM estimates. In case the HR confidence interval is not mentioned in the original article, the results in Figures A1, A2, and A3 suggest that only using HR yields Bayes factors that are reasonable approximations to the true Bayes factor (Appendix B shows a follow-up analysis of the influence of specific weight combinations for HR, $\mathrm{CI}\,(\mathrm{HR})_{LB}$, and $\mathrm{CI}\,(\mathrm{HR})_{UB}$ on the accuracy of the simulated Bayes factors, demonstrating that specific choices of weights do not matter). 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.

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) and computed 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. Here as well, we used a Normal prior with a mean of $\mu = 0$ and a standard deviation of $\sigma = 1$ for the β parameter.

The results are shown in Appendix C. Figure C1 indicates that a small number of PSO iterations is required when only using HR because the variance of the Bayes factors does not improve markedly when more than approximately 30 or 100 iterations are used. In contrast, Figure C2 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.

Reanalysis of [5]

To demonstrate our proposed procedure for simulating data from summary statistics and computing one Bayes factor for each simulated data set, we again reanalyzed the study reported in [5]. But this time we will only used the summary statistics reported in their article. As in the previous section, our reanalysis of [5] omitted the stratification by actual disease severity and we used Efron's instead of Breslow's method. 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 [5] (i.e., HR = 1.29, 95% CI = [1.12, 1.49]). We used our computed summary statistics as if they were the only results provided in [5].

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
)
```

Subsequently, we computed one Bayes factor for each of the 100 simulated data sets. We used a truncated (because we have a one-sided $\mathcal{H}_1$) Normal prior for β 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
    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
)
```

A histogram of the resulting Bayes factors can be found in Figure 1. The simulated Bayes factors range between $\mathrm{BF}_{+0} = 133.6$ and $\mathrm{BF}_{+0} = 135.7$, all providing 'decisive'[43], 'extreme'[50], or 'strong'[45] evidence supporting the conclusion of [5] 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 [5]. 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).

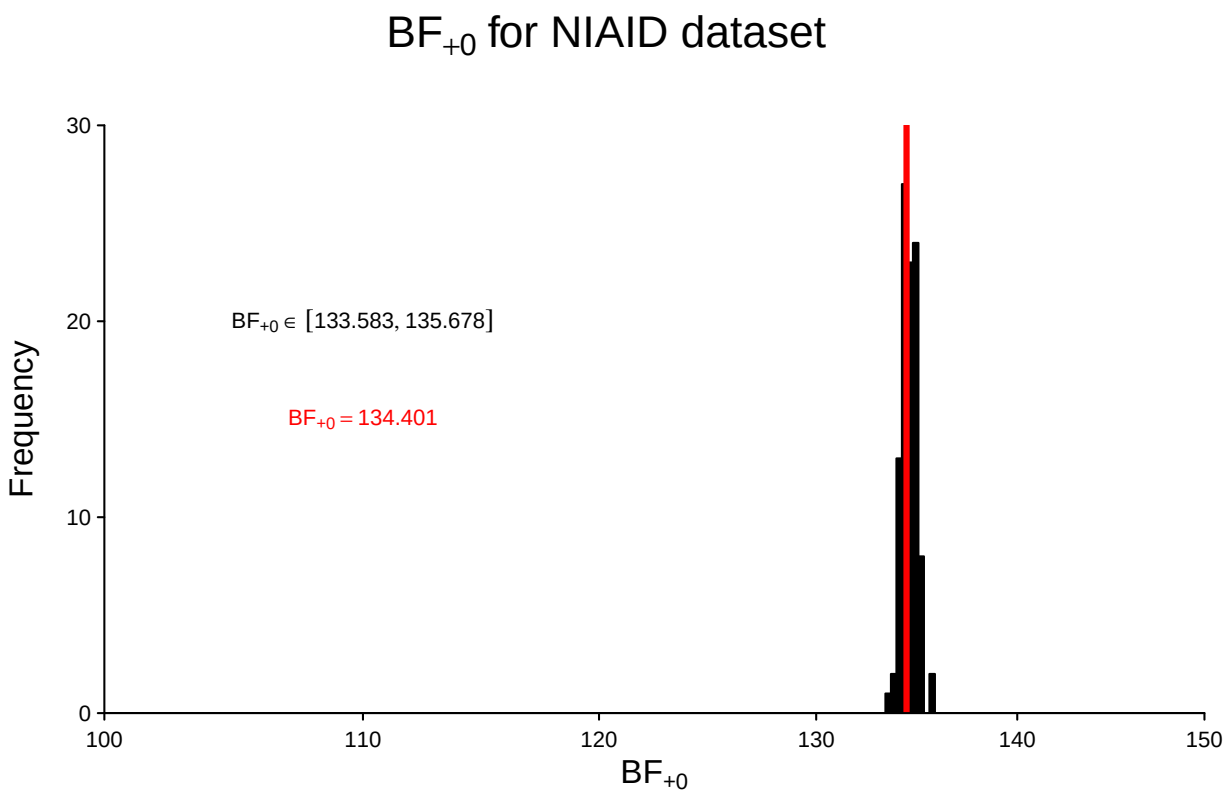

Figure 1

Distribution of BF_{+0} applied to 100 *simulated data sets for the data set described in [5]. 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.*

Comparison with Savage-Dickey Normal Approximation

In the following, we compare our procedure for computing Bayes factors with an alternative procedure that tries to approximate Bayes factors.

[4] introduced a generic method that uses a Normal approximation of the likelihood function to compute a Bayes factor for various statistical designs. If $\mathcal{H}_0$ is a point 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 β, evaluated at the null value β_0. This ratio is called the Savage-Dickey density ratio [e.g., 25]:

$$\mathrm{BF}_{10} = \frac{f\left(\beta = \beta_0 \mid \mathcal{H}_1\right)}{f\left(\beta = \beta_0 \mid D,\ \mathcal{H}_1\right)}. \tag{14}$$

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 μ and variance σ^2, a closed-form solution for the Bayes factor is available [cf. the last equation on p. 3 of 4]:

$$\mathrm{BF}_{01} = \sqrt{\frac{\sigma^2 + SE\left(\hat{\beta}\right)^2}{SE\left(\hat{\beta}\right)^2}} \exp\left(-\frac{1}{2}\left[\frac{\left(\hat{\beta} - \beta_0\right)^2}{SE\left(\hat{\beta}\right)^2} - \frac{\left(\hat{\beta} - \mu\right)^2}{\sigma^2 + SE\left(\hat{\beta}\right)^2}\right]\right). \tag{15}$$

[4] 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 [4] also yields accurate Bayes factors for semi-parametric Cox models and how they compare to the Bayes factors resulting from our method. This is important because the method by [4] provides a closed-form solution for computing 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 [4] would be preferable. We refer the reader to Appendix D for a demonstration of typical run times of our method. We took the three example data sets mentioned before and computed one Bayes factor through our method and one Bayes factor through the method by [4], 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 computed one Bayes factor for each simulated data set using our method.

The results are shown in Figure 2, where the red vertical line represents the Bayes factor for the full data set computed using our method, the blue vertical line represents the Bayes factor resulting from the Savage-Dickey Normal approximation method advocated in [4], and the histogram shows Bayes factors from our method when using summary statistics. The Bayes factors resulting from the method by [4] 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 2, our Bayes factors are much more accurate (i.e., the distance between histograms and red lines is less than the distance between the blue lines and red lines). 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 [4] (see Appendix D).

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 [23] 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 compute Bayes factors for simple Cox models and offered the R package "baymedr" [52] as an easy-to-use implementation. "baymedr" can be used to compute a Bayes factor for full data and to simulate multiple Bayes factors based on summary statistics as reported in articles.

Our procedure for computing 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

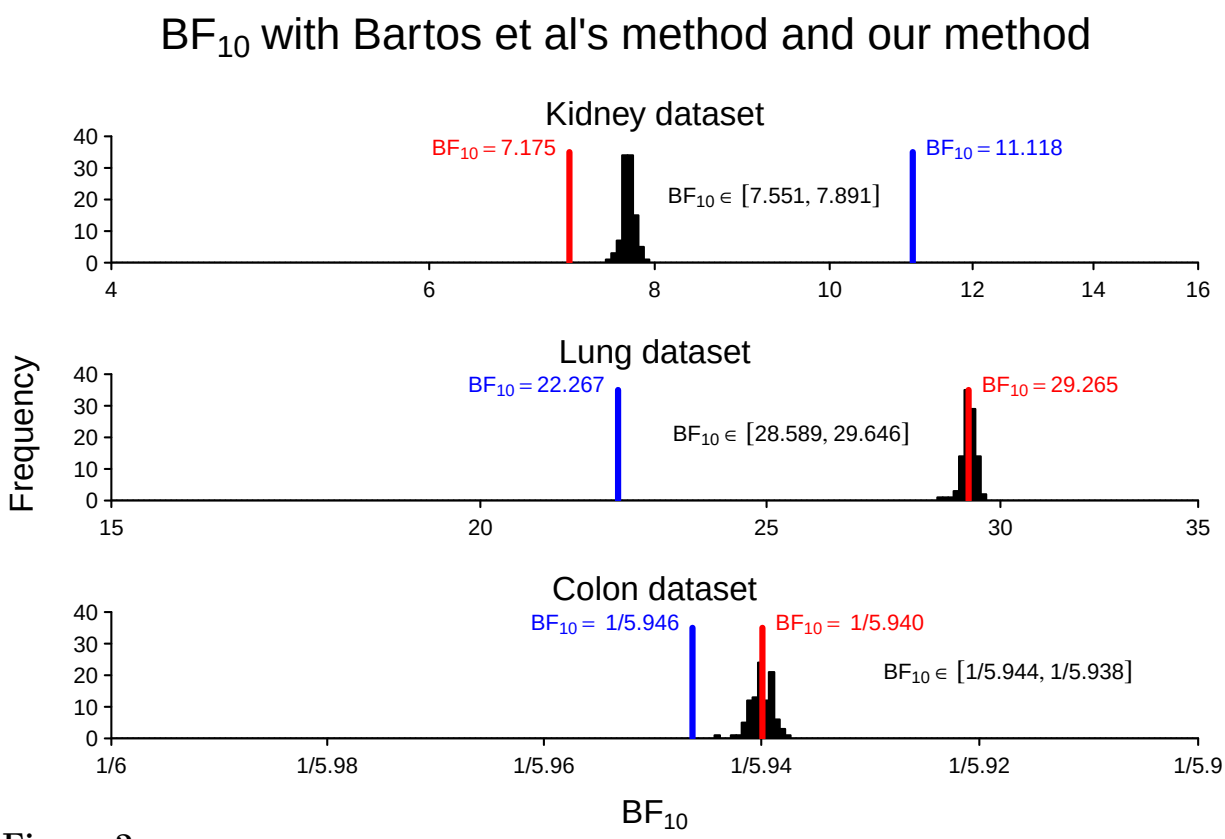

Figure 2

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 [4].*

effect is present and evidence than an effect is absent and allows for optional stopping during data collection.

At the same time, many features and functionalities are still missing. For example, it would be desirable to also make the Cox partial likelihood [23] and Breslow's approximation to the true partial likelihood [13] available. Moreover, the procedure implemented in "baymedr" should allow researchers to compute 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 compute 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., 10, 14, 32, 71]; the posterior samples could then be used to estimate the marginal likelihood through bridge sampling [e.g., 36].

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 computing 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" [52], an R package that is aimed at all researchers desiring to compute a Bayes factor for their Cox regression.

Acknowledgments

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.

References

[1] P. Armitage, C. K. McPherson, and B. C. Rowe. "Repeated Significance Tests on Accumulating Data". In: *Journal of the Royal Statistical Society. Series A (General)* 132.2 (1969), pp. 235–244. DOI: 10.2307/2343787.

[2] David Bakan. "The Test of Significance in Psychological Research". In: *Psychological Bulletin* 66.6 (1966), pp. 423–437. DOI: 10.1037/h0020412.

[3] František Bartoš. *RoBSA: An R Package for Robust Bayesian Survival Analyses*. R package version 1.0.0. 2022. URL: https://CRAN.R-project.org/package=RoBSA.

[4] František Bartoš and Eric-Jan Wagenmakers. "A General Approximation to Nested Bayes Factors with Informed Priors". In: *Stat* 12.1 (2023), e600. DOI: 10.1002/sta4.600.

[5] John H. Beigel et al. "Remdesivir for the Treatment of Covid-19 - Final Report". In: *The New England Journal of Medicine* 383.19 (2020), pp. 1813–1826. DOI: 10.1056/NEJMoa2007764.

[6] Dmitry I. Belov and Ronald D. Armstrong. "Distributions of the Kullback–Leibler Divergence with Applications". In: *British Journal of Mathematical and Statistical Psychology* 64.2 (2011), pp. 291–309. DOI: https://doi.org/10.1348/000711010X522227.

[7] Claus Bendtsen. *pso: Particle Swarm Optimization*. R package version 1.0.4. 2022. URL: https://CRAN.R-project.org/package=pso.

[8] James O. Berger and Mohan Delampady. "Testing Precise Hypotheses". In: *Statistical Science* 2.3 (1987), pp. 317–335. DOI: 10.1214/ss/1177013238.

[9] James O. Berger and Thomas Sellke. "Testing a Point Null Hypothesis: The Irreconcilability of p Values and Evidence". In: *Journal of the American Statistical Association* 82.397 (1987), pp. 112–122. DOI: 10.2307/2289131.

[10] Michael Betancourt. *A Conceptual Introduction to Hamiltonian Monte Carlo*. 2018. arXiv: 1701.02434v2 [stat.ME].

[11] M. J. Bradburn et al. "Survival Analysis Part II: Multivariate Data Analysis - An Introduction to Concepts and Methods". In: *British Journal of Cancer* 89.3 (2003), pp. 431–436. DOI: 10.1038/sj.bjc.6601119.

[12] M. J. Bradburn et al. "Survival Analysis Part III: Multivariate Data Analysis - Choosing a Model and Assessing Its Adequacy and Fit". In: *British Journal of Cancer* 89.4 (2003), pp. 605–611. DOI: 10.1038/sj.bjc.6601120.

[13] Norman Breslow. "Covariance Analysis of Censored Survival Data". In: *Biometrics* 30.1 (1974), pp. 89–99. DOI: 10.2307/2529620.

[14] Steven Brooks et al. *Handbook of Markov Chain Monte Carlo*. New York, NY: Chapman & Hall/CRC, 2011.

[15] Iain Chalmers and Paul Glasziou. "Avoidable Waste in the Production and Reporting of Research Evidence". In: *The Lancet* 374.9683 (2009), pp. 86–89. DOI: 10.1016/S0140-6736(09)60329-9.

[16] David Chavalarias et al. "Evolution of Reporting p Values in the Biomedical Literature, 1990-2015". In: *Journal of the American Medical Association* 315.11 (2016), pp. 1141–1148. DOI: 10.1001/jama.2016.1952.

[17] Erik Christensen. "Methodology of Superiority vs. Equivalence Trials and Non-Inferiority Trials". In: *Journal of Hepatology* 46.5 (2007), pp. 947–954. DOI: 10.1016/j.jhep.2007.02.015.

[18] T. G. Clark et al. "Survival Analysis Part I: Basic Concepts and First Analyses". In: *British Journal of Cancer* 89.2 (2003), pp. 232–238. DOI: 10.1038/sj.bjc.6601118.

[19] T. G. Clark et al. "Survival Analysis Part IV: Further Concepts and Methods in Survival Analysis". In: *British Journal of Cancer* 89.5 (2003), pp. 781–786. DOI: 10.1038/sj.bjc.6601117.

[20] Maurice Clerc. *Particle Swarm Optimization*. London, UK: ISTE, 2006.

[21] Jacob Cohen. "The Earth Is Round ($p < .05$)". In: *American Psychologist* 49.12 (1994), pp. 997–1003. DOI: 10.1037/0003-066X.49.12.997.

[22] David Collett. *Modelling Survival Data in Medical Research*. 3rd. Boca Raton, FL: CRC Press, 2015.

[23] David R. Cox. "Regression Models and Life-tables". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2 (1972), pp. 187–202. DOI: 10.1111/j.2517-6161.1972.tb00899.x.

[24] Nathan C. Dean et al. "Decreased Mortality After Implementation of a Treatment Guideline for Community-acquired Pneumonia". In: *The American Journal of Medicine* 110.6 (2001), pp. 451–457. DOI: 10.1016/S0002-9343(00)00744-0.

[25] James M. Dickey and Bennet P. Lientz. "The Weighted Likelihood Ratio, Sharp Hypotheses about Chances, the Order of a Markov Chain". In: *The Annals of Mathematical Statistics* 41.1 (1970), pp. 214–226. DOI: 10.1214/aoms/1177697203.

[26] Zoltan Dienes. "Bayesian Versus Orthodox Statistics: Which Side Are You on?" In: *Perspectives on Psychological Science* 6.3 (2011), pp. 274–290. DOI: 10.1177/1745691611406920.

[27] Bradley Efron. "The Efficiency of Cox's Likelihood Function for Censored Data". In: *Journal of the American Statistical Association* 72.359 (1977), pp. 557–565. DOI: 10.1080/01621459.1977.10480613.

[28] Andy P. Field and Raphael Gillett. "How to Do a Meta-analysis". In: *British Journal of Mathematical and Statistical Psychology* 63 (2010), pp. 665–694. DOI: 10.1348/000711010X502733.

[29] Lawrence M. Friedman et al. *Fundamentals of Clinical Trials*. 4th. New York, NY: Springer, 2010.

[30] C. R. Gallistel. "The Importance of Proving the Null". In: *Psychological Review* 116.2 (2009), pp. 439–453. DOI: 10.1037/a0015251.

[31] Gerd Gigerenzer. "Mindless Statistics". In: *The Journal of Socio-Economics* 33.5 (2004), pp. 587–606. DOI: 10.1016/j.socec.2004.09.033.

[32] W. R. Gilks, S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Boca Raton, FL: Chapman & Hall/CRC, 1995.

[33] Steven N. Goodman. "Toward Evidence-based Medical Statistics. 1: The p Value Fallacy". In: *Annals of Internal Medicine* 130.12 (1999), pp. 995–1004. DOI: 10.7326/0003-4819-130-12-199906150-00008.

[34] Steven N. Goodman. "Toward Evidence-based Medical Statistics. 2: The Bayes Factor". In: *Annals of Internal Medicine* 130.12 (1999), pp. 1005–1013. DOI: 10.7326/0003-4819-130-12-199906150-00019.

[35] Steven N. Goodman. "A Dirty Dozen: Twelve p-value Misconceptions". In: *Seminars in Hematology* 45.3 (2008), pp. 135–140. DOI: 10.1053/j.seminhematol.2008.04.003.

[36] Quentin F. Gronau et al. "A Tutorial on Bridge Sampling". In: *Journal of Mathematical Psychology* 81 (2017), pp. 80–97. DOI: 10.1016/j.jmp.2017.09.005.

[37] Frank E. Harrell. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. 2nd. New York, NY: Springer, 2015.

[38] Frank E. Harrell. *rms: Regression Modeling Strategies*. R package version 6.3-0. 2022. URL: https://CRAN.R-project.org/package=rms.

[39] David W. Hosmer, Stanley Lemeshow, and Susanne May. *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*. 2nd. Hoboken, NJ: John Wiley & Sons, 2008.

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

[41] Harold Jeffreys. *Theory of Probability*. Oxford, UK: The Clarendon Press, 1939.

[42] Harold Jeffreys. *Theory of Probability*. 2nd. Oxford, UK: The Clarendon Press, 1948.

[43] Harold Jeffreys. *Theory of Probability*. 3rd. Oxford, UK: Oxford University Press, 1961.

[44] Edward L. Kaplan and Paul Meier. "Nonparametric Estimation from Incomplete Observations". In: *Journal of the American Statistical Association* 53.282 (1958), pp. 457–481. DOI: 10.1080/01621459.1958.10501452.

[45] Robert E. Kass and Adrian E. Raftery. "Bayes Factors". In: *Journal of the American Statistical Association* 90.430 (1995), pp. 773–795. DOI: 10.2307/2291091.

[46] James Kennedy and Russell Eberhart. "Particle Swarm Optimization". In: *Proceedings of ICNN'95 - International Conference on Neural Networks*. Vol. 4. IEEE. 1995, pp. 1942–1948. DOI: 10.1109/ICNN.1995.488968.

[47] Christian Keysers, Valeria Gazzola, and Eric-Jan Wagenmakers. "Using Bayes Factor Hypothesis Testing in Neuroscience to Establish Evidence of Absence". In: *Nature Neuroscience* 23 (2020), pp. 788–799. DOI: 10.1038/s41593-020-0660-4.

[48] John P. Klein and Melvin L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. New York, NY: Springer, 1997.

[49] J. A. Laurie et al. "Surgical Adjuvant Therapy of Large-bowel Carcinoma: An Evaluation of Levamisole and the Combination of Levamisole and Fluorouracil. The North Central Cancer Treatment Group and the Mayo Clinic". In: *Journal of Clinical Oncology* 7.10 (1989), pp. 1447–1456. DOI: 10.1200/JCO.1989.7.10.1447.

[50] Michael D. Lee and Eric-Jan Wagenmakers. *Bayesian Cognitive Modeling: A Practical Course*. Cambridge, UK: Cambridge University Press, 2013.

[51] Kwan-Moon Leung, Robert M. Elashoff, and Abdelmonem A. Afifi. "Censoring Issues in Survival Analysis". In: *Annual Review of Public Health* 18 (1997), pp. 83–104. DOI: 10.1146/annurev.publhealth.18.1.83.

[52] Maximilian Linde, Don van Ravenzwaaij, and Jorge N. Tendeiro. *baymedr: Computation of Bayes Factors for Common Biomedical Designs*. R package version 0.1.1.9000. 2022. URL: https://github.com/maxlinde/baymedr.

[53] C. L. Loprinzi et al. "Prospective Evaluation of Prognostic Variables from Patient-completed Questionnaires. North Central Cancer Treatment Group". In: *Journal of Clinical Oncology* 12.3 (1994), pp. 601–607. DOI: 10.1200/JCO.1994.12.3.601.

[54] C. A. McGilchrist and C. W. Aisbett. "Regression with Frailty in Survival Analysis". In: *Biometrics* 47.2 (1991), pp. 461–466. DOI: 10.2307/2532138.

[55] Blakeley B. McShane et al. "Abandon Statistical Significance". In: *The American Statistician* 73.sup1 (2019), pp. 235–245. DOI: 10.1080/00031305.2018.1527253.

[56] Richard D. Morey and Jeffrey N. Rouder. *BayesFactor: Computation of Bayes Factors for Common Designs*. R package version 0.9.12-4.2. 2018. URL: https://CRAN.R-project.org/package=BayesFactor.

[57] Rupert M. Pearse et al. "Mortality After Surgery in Europe: A 7 Day Cohort Study". In: *The Lancet* 380.9847 (2012), pp. 1059–1065. DOI: 10.1016/S0140-6736(12)61148-9.

[58] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2022. URL: https://www.R-project.org/.

[59] Jeffrey N. Rouder. "Optional Stopping: No Problem for Bayesians". In: *Psychonomic Bulletin & Review* 21.2 (2014), pp. 301–308. DOI: 10.3758/s13423-014-0595-4.

[60] Jeffrey N. Rouder et al. "Bayesian t Tests for Accepting and Rejecting the Null Hypothesis". In: *Psychonomic Bulletin & Review* 16.2 (2009), pp. 225–237. DOI: 10.3758/PBR.16.2.225.

[61] Adam N. Sanborn and Thomas T. Hills. "The Frequentist Implications of Optional Stopping on Bayesian Hypothesis Tests". In: *Psychonomic Bulletin & Review* 21.2 (2014), pp. 283–300. DOI: 10.3758/s13423-013-0518-9.

[62] Felix D. Schönbrodt and Eric-Jan Wagenmakers. "Bayes Factor Design Analysis: Planning for Compelling Evidence". In: *Psychonomic Bulletin and Review* 25.1 (2018), pp. 128–142. DOI: 10.3758/s13423-017-1230-y.

[63] Felix D. Schönbrodt et al. "Sequential Hypothesis Testing with Bayes Factors: Efficiently Testing Mean Differences". In: *Psychological Methods* 22.2 (2017), pp. 322–339. DOI: 10.1037/met0000061.

[64] Stephen Senn. *Statistical Issues in Drug Development*. 2nd. Chichester, UK: John Wiley & Sons, 2008.

[65] Yuhui Shi and Russell Eberhart. "A Modified Particle Swarm Optimizer". In: *1998 IEEE International Conference on Evolutionary Computation Proceedings. IEEE World Congress on Computational Intelligence*. 1998, pp. 69–73. DOI: 10.1109/ICEC.1998.699146.

[66] Sandip Sinharay and Hal S. Stern. "On the Sensitivity of Bayes Factors to the Prior Distributions". In: *The American Statistician* 56.3 (2002), pp. 196–201. DOI: 10.1198/000313002137.

[67] Alex J. Sutton and R. Abrams Keith. "Bayesian Methods in Meta-analysis and Evidence Synthesis". In: *Statistical Methods in Medical Research* 10.4 (2001), pp. 277–303. DOI: 10.1177/096228020101000404.

[68] Jorge N. Tendeiro, Henk A. L. Kiers, and Don Van Ravenzwaaij. "Worked-out Examples of the Adequacy of Bayesian Optional Stopping". In: *Psychonomic Bulletin & Review* 29.1 (2022), pp. 70–87. DOI: 10.3758/s13423-021-01962-5.

[69] Terry M. Therneau. *A Package for Survival Analysis in R*. R package version 3.2-13. 2021. URL: https://CRAN.R-project.org/package=survival.

[70] Terry M. Therneau and Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. New York, NY: Springer, 2000.

[71] Don van Ravenzwaaij, Pete Cassey, and Scott D. Brown. "A Simple Introduction to Markov Chain Monte–Carlo Sampling". In: *Psychonomic Bulletin & Review* 25.1 (2018), pp. 143–154. DOI: 10.3758/s13423-016-1015-8.

[72] Don van Ravenzwaaij and John P. A. Ioannidis. "A Simulation Study of the Strength of Evidence in the Recommendation of Medications Based on Two Trials with Statistically Significant Results". In: *PLoS ONE* 12.3 (2017), e0173184. DOI: 10.1371/journal.pone.0173184.

[73] Don van Ravenzwaaij et al. "Bayes Factors for Superiority, Non-inferiority, and Equivalence Designs". In: *BMC Medical Research Methodology* 19.1 (2019), p. 71. DOI: 10.1186/s12874-019-0699-7.

[74] Wolf Vanpaemel. "Prior Sensitivity in Theory Testing: An Apologia for the Bayes Factor". In: *Journal of Mathematical Psychology* 54.6 (2010), pp. 491–498. DOI: 10.1016/j.jmp.2010.07.003.

[75] Eric-Jan Wagenmakers. "A Practical Solution to the Pervasive Problems of p Values". In: *Psychonomic Bulletin & Review* 14.5 (2007), pp. 779–804. DOI: 10.3758/BF03194105.

[76] Eric-Jan Wagenmakers et al. "Bayesian Inference for Psychology. Part I: Theoretical Advantages and Practical Ramifications". In: *Psychonomic Bulletin & Review* 25.1 (2018), pp. 35–57. DOI: 10.3758/s13423-017-1343-3.

[77] Ronald L. Wasserstein and Nicole A. Lazar. "The ASA's Statement on p-values: Context, Process, and Purpose". In: *The American Statistician* 70.2 (2016), pp. 129–133. DOI: 10.1080/00031305.2016.1154108.

[78] Hadley Wickham, Jim Hester, and Winston Chang. *devtools: Tools to Make Developing R Packages Easier*. R package version 2.2.0. 2019. URL: https://CRAN.R-project.org/package=devtools.

Appendix A

Results for Weights of Summary Statistics

Figures A1, A2, and A1 show the bias and variance of the simulated Bayes factors based on different combinations of weights of summary statistics in the computation of the loss function for the Kidney [54], Lung [53], and Colon [49] data sets, respectively, from the "survival" R package [69]. See main text for details.

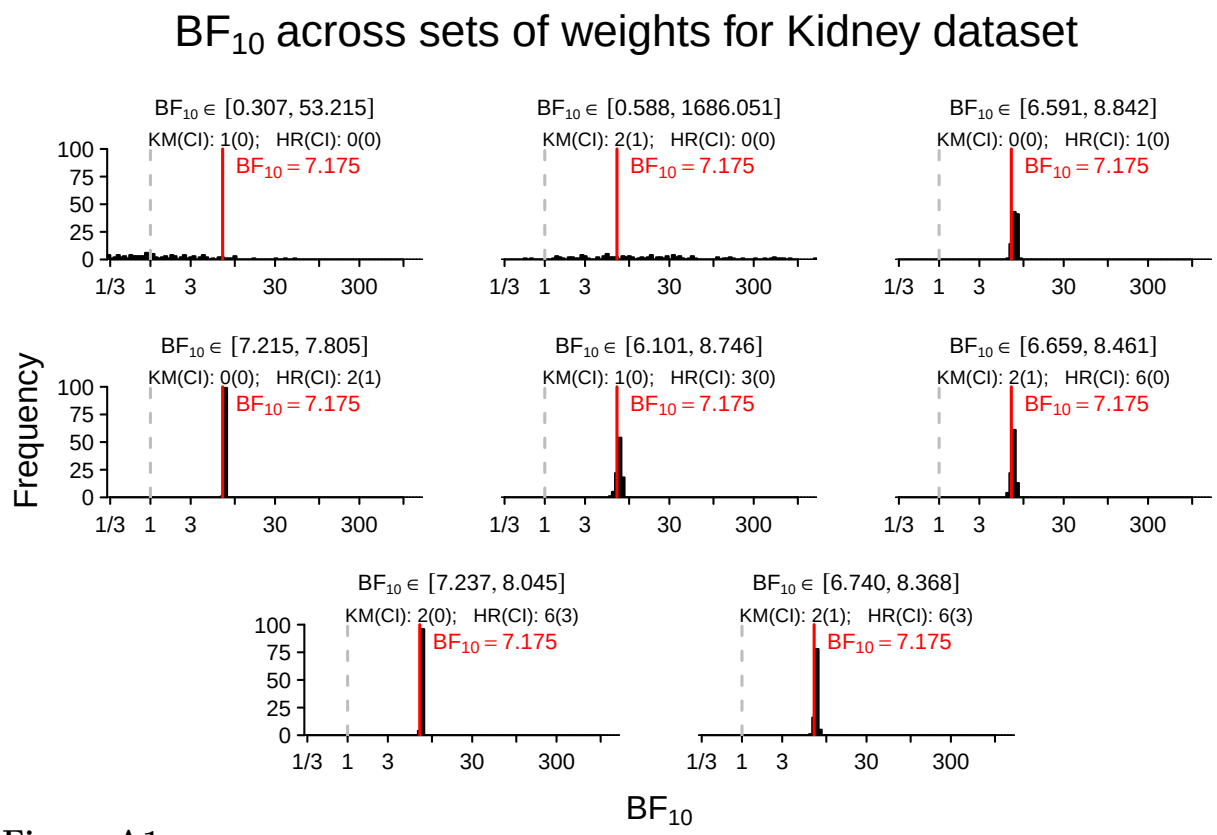

Figure A1

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.

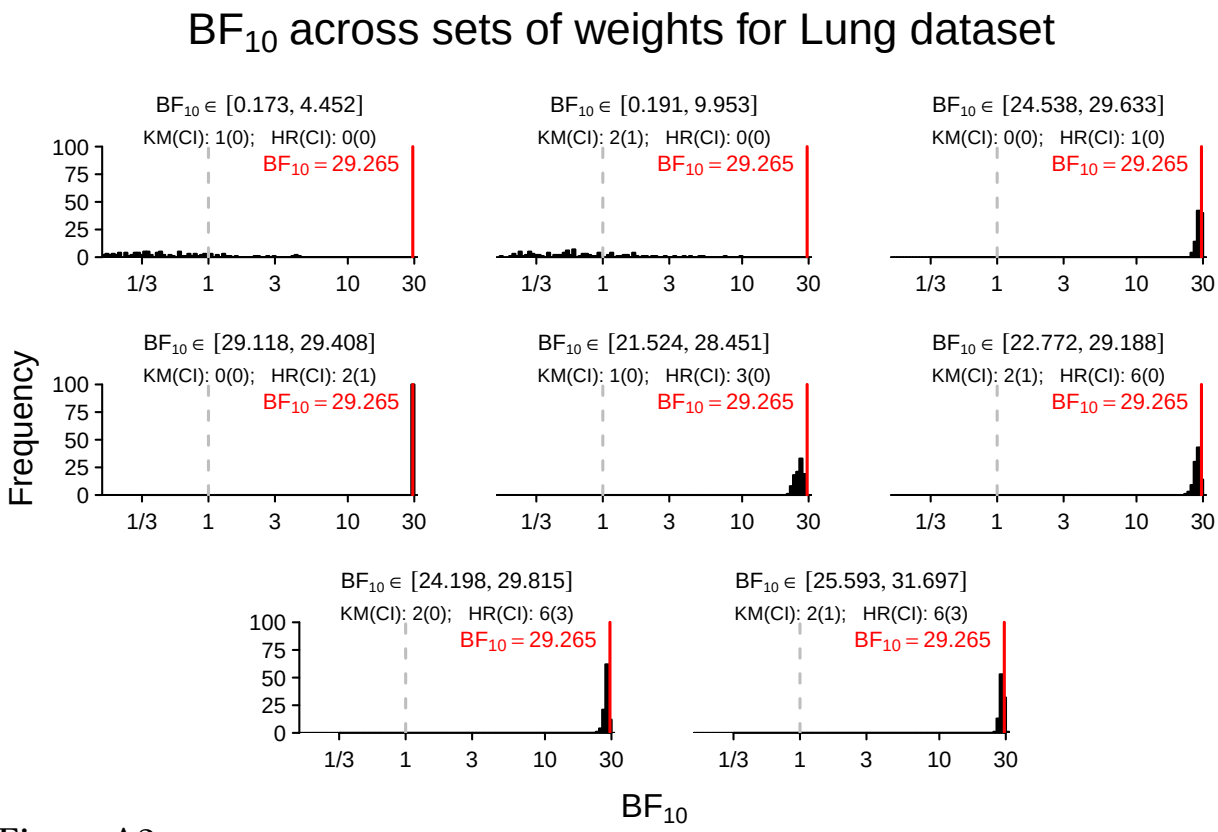

Figure A2

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.

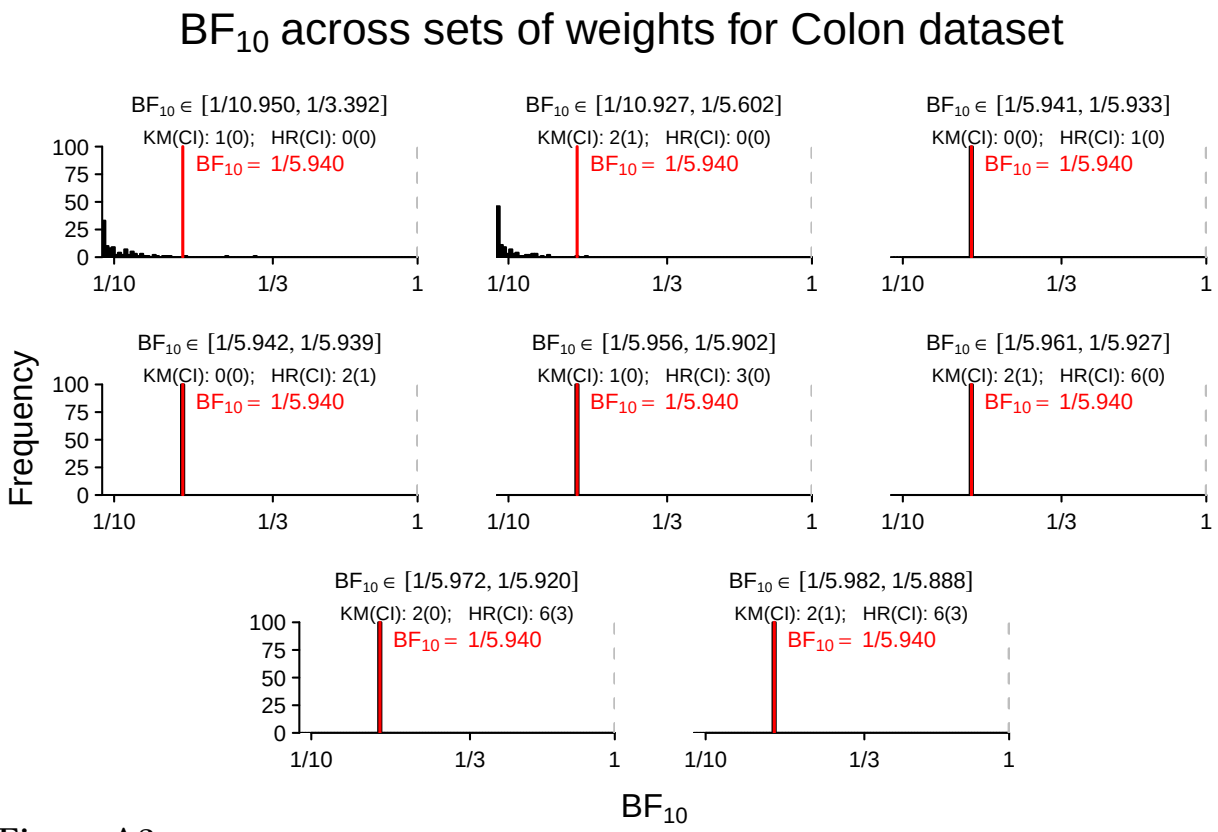

Figure A3

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.*

Appendix B

Follow-up Analyses on the Influence of Different Weight Combinations

In the main manuscript, we have seen through simulations that using HR, $\mathrm{CI}\,(\mathrm{HR})_{LB}$, and $\mathrm{CI}\,(\mathrm{HR})_{UB}$ as summary statistics yields the most accurate Bayes factors. We conducted follow-up simulations to investigate whether it is relevant which weights for HR, $\mathrm{CI}\,(\mathrm{HR})_{LB}$, and $\mathrm{CI}\,(\mathrm{HR})_{UB}$ are chosen. Figure B1 shows the Bayes factors for 10 different weight settings for the Kidney, Lung, and Colon data sets. For each weight setting, 20 data sets were simulated with 200 PSO iterations and, accordingly, 20 Bayes factors were computed. It can be seen that it does not make a difference which specific weights are chosen, except in the situation where only HR is available, which diminishes the accuracy of the simulated Bayes factors.

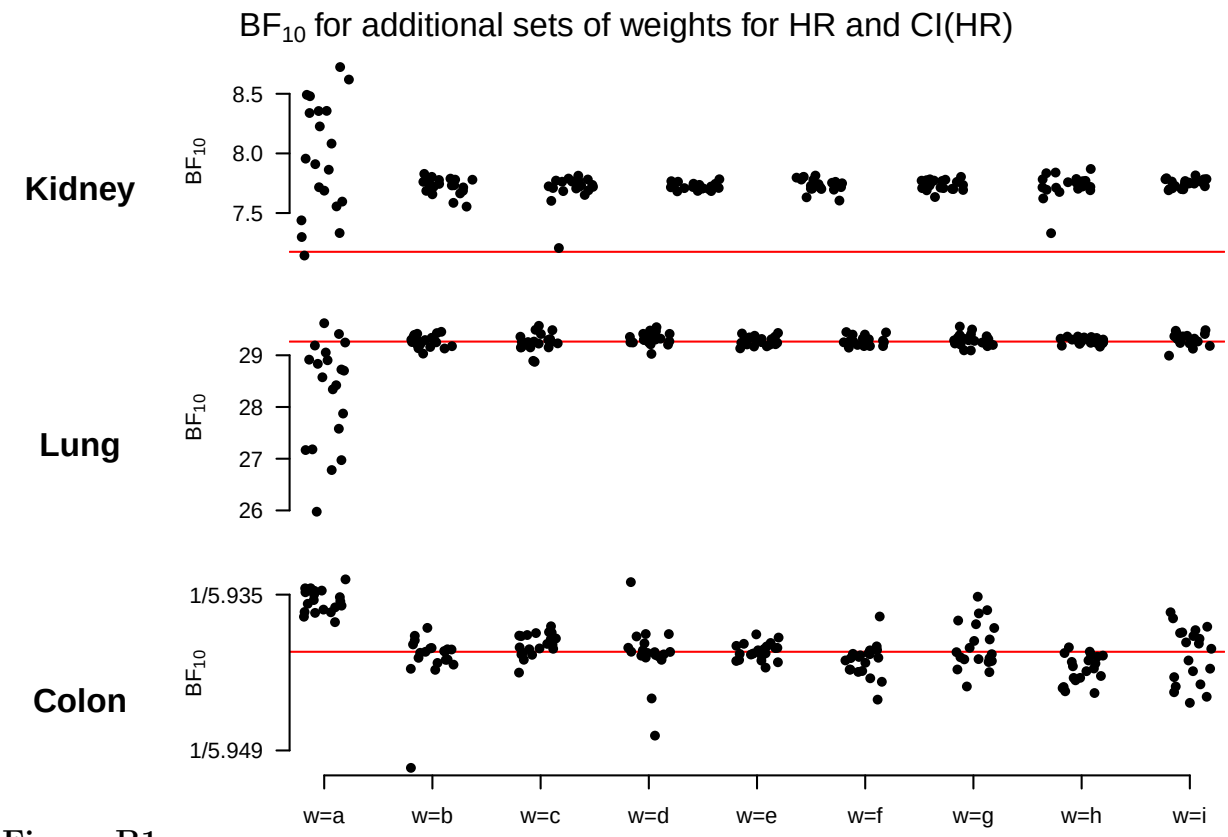

Figure B1

Accuracy of simulated Bayes factors as a function of weight combinations. The red horizontal line displays the true Bayes factor. The x-axis labels, matching the pattern $HR[CI(HR)_{LB}, CI(HR)_{UB}]$*, are: a=1[0, 0], b=2[1, 1], c=4[1, 1], d=1[2, 2], e=1[4, 4], f=3[1, 2], g=3[2, 1], h=1[2, 3], i=1[3, 2].*

Appendix C

Results for Maximum Number of PSO Iterations

Figures C1, and C2 show the bias and variance of the simulated Bayes factors based on different maximum numbers of PSO iterations for the Kidney [54], Lung [53], and Colon [49] data sets from the "survival" R package [69], when using HR or a combination of HR and CI (HR), respectively. See main text for details.

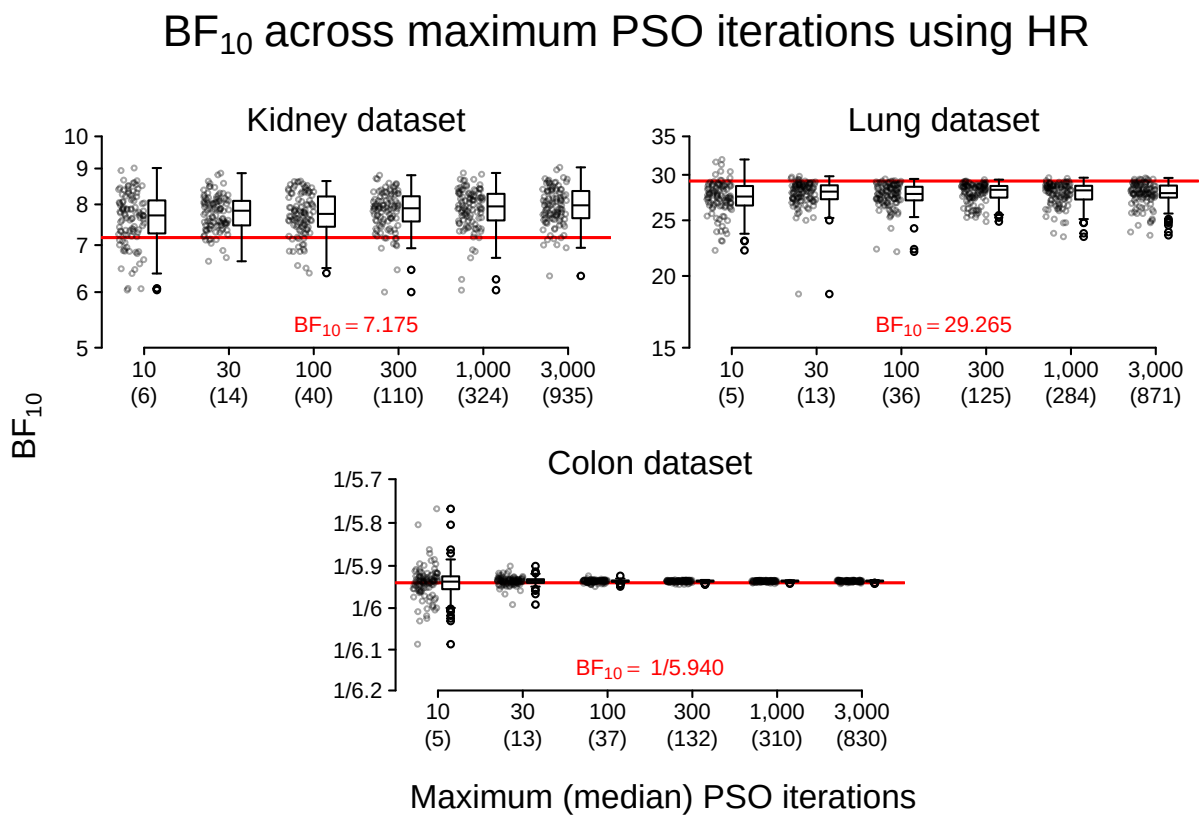

Figure C1

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.*

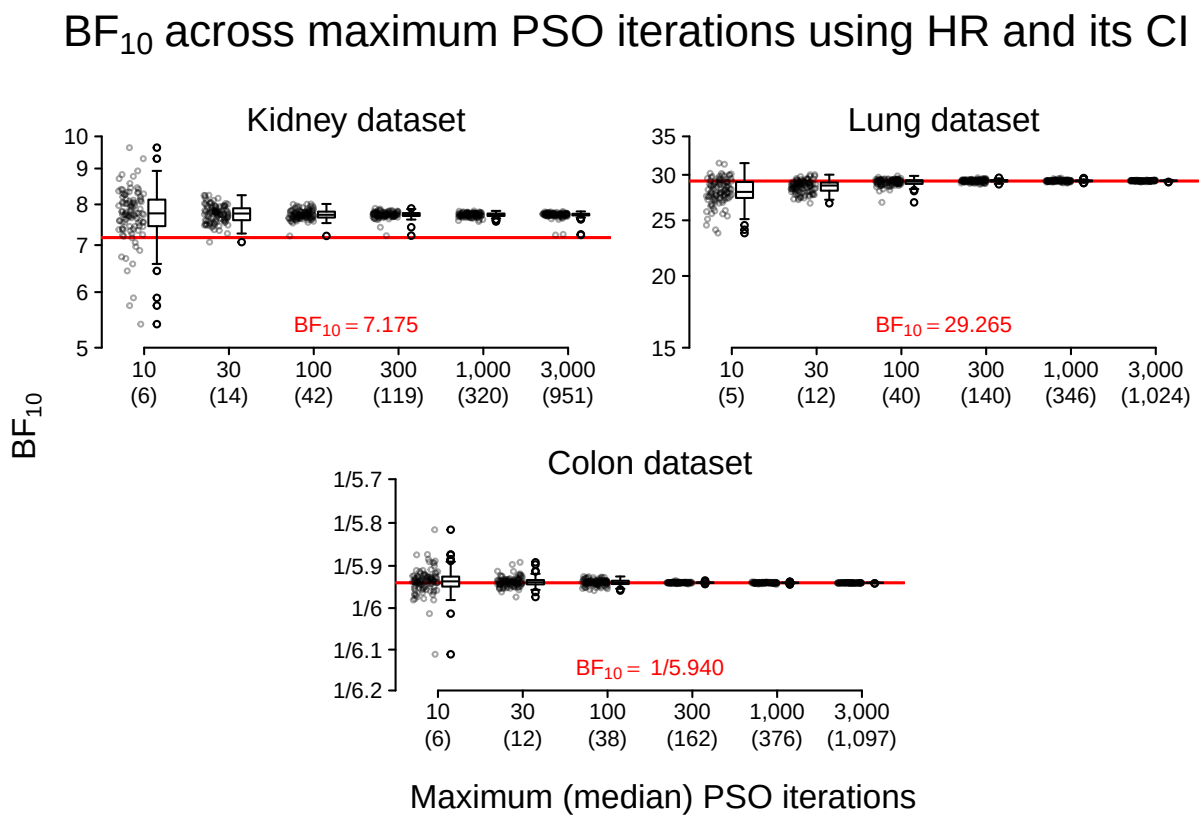

Figure C2

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.*

Appendix D

Run Time of Data Simulation and Bayes Factor Computation

Figure D1 shows the run time of our proposed procedure, where the process of data simulation is shown in the left panel and Bayes factor computation in the right panel. The run time was measured on a typical Windows 10 ($x64$) computer, with 16GB of RAM and an 11th Gen Intel(R) Core(TM) i7-1165G7 @ 2.80GHz CPU with 8 cores. For the measurement of the run times, we used the Kidney [54], Lung [53], and Colon [49] data sets that are available through the "survival" R package [69]. Furthermore, for data simulation, we either used 1, 2, or 4 CPU cores. For all data simulations, we simulated 20 data sets with exactly 100 PSO iterations. Both data simulation and computing Bayes factors were repeated 40 times to obtain a distribution of run times.

Figure D1 indicates that data simulation takes longer than computing Bayes factors. Using our settings, simulation of 20 data sets with 100 PSO iterations took approximately $60 - 150$s, $70 - 180$s, and $110 - 300$s for the Kidney, Lung, and Colon data sets, respectively, using $1 - 4$ CPU cores. Computing Bayes factors took approximately 4, 18, and 90s for the Kidney, Lung, and Colon data sets, respectively. It can be seen that both data simulation and Bayes factor computation take longer the larger the sample size of the underlying data set is. Moreover, it is visible that the run time for data simulation can be drastically reduced by using multiple CPU cores. However, this is only possible when more than one data set is to be simulated. Ideally, the number of data sets to be simulated is a multiple of the number of used CPU cores. Further, parallel processing with multiple CPU cores is only beneficial when a considerable number of PSO iterations is used and when the sample size of the underlying data set is not too small. Otherwise, the initialization of the CPU cores takes up too much time for the benefits of parallel processing to unfold.

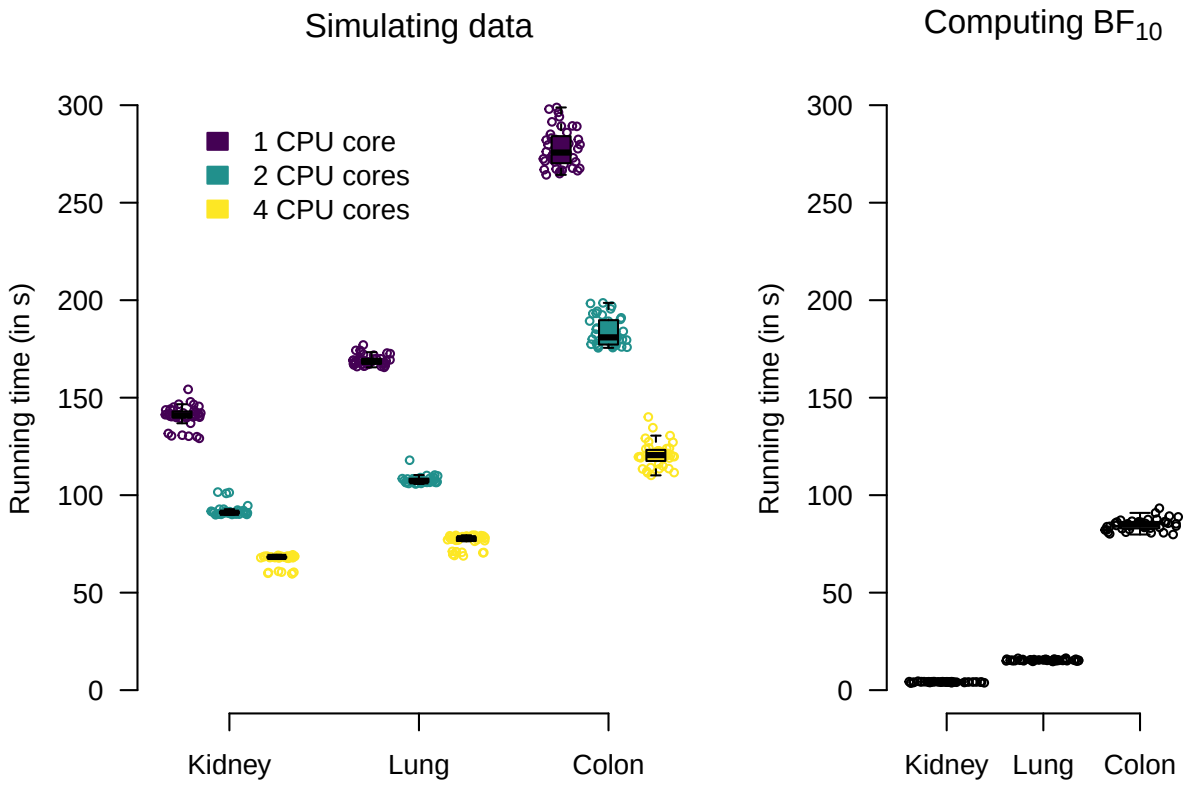

Figure D1

Runtime for data simulation of 20 *data sets (left panel) and for Bayes factor computation (right panel) using the Kidney, Lung, and Colon data sets. For each data set, data simulation and Bayes factor computation was repeated* 40 *times. For data simulation (left panel), run times with* 1, 2, *and* 4 *CPU cores are shown in purple, green, and yellow, respectively.*

Appendix E

Exploration of Factors Contributing to the Accuracy Difference Between Our Method and the Method by [4]

In the main manuscript we found that our method for simulating Bayes factors provides more accurate Bayes factors than the method by [4] (see Figure 2). To explore possible reasons for this discrepancy, we conducted further simulations.

We took the Kidney, Lung, and Colon data sets and generated data sets of variable sizes through bootstrapping (i.e., sampling with replacement). The sizes of the bootstrapped data sets were $n = \{40, 80, 160, 320, 640, 1280\}$. This process was repeated 10 times. Thus, in total we have $3 \times 6 \times 10 = 180$ bootstrapped data sets. Subsequently, two separate analyses were conducted.

First, for each bootstrapped data set, we used the full data to compute a Bayes factor using our method. This Bayes factor served as the "truth". Moreover, we applied our approach and the approach by [4] on each bootstrapped data set. In other words, we have (1) the "true" Bayes factor, (2) the median Bayes factor resulting from our approach (median because we simulate 8 data sets), and (3) the Bayes factor resulting from the method by [4] for each bootstrapped data set. Then, for each bootstrapped data set we computed two numbers: 1) $log\,(BF_{\text{Our}}/BF_{\text{True}})$ and 2) $log\,(BF_{\text{Bartos}}/BF_{\text{True}})$. Figure E1 shows these accuracy ratios as a function of the sample size of the bootstrapped data sets. For the Kidney and Lung data sets, the accuracy of the approach by [4] decreases as the sample size increases. For the Colon data set, accuracy is almost equal across sample sizes for the approach by [4], except for small sample sizes, where less accuracy is observed. The accuracy of the Bayes factors resulting from the summary BF procedure are good for all data sets and for almost all sample sizes. Only for the Kidney data set it was observed that accuracy diminishes for higher sample sizes.

Second, we explored whether the shape of the likelihood influences the accuracy of the

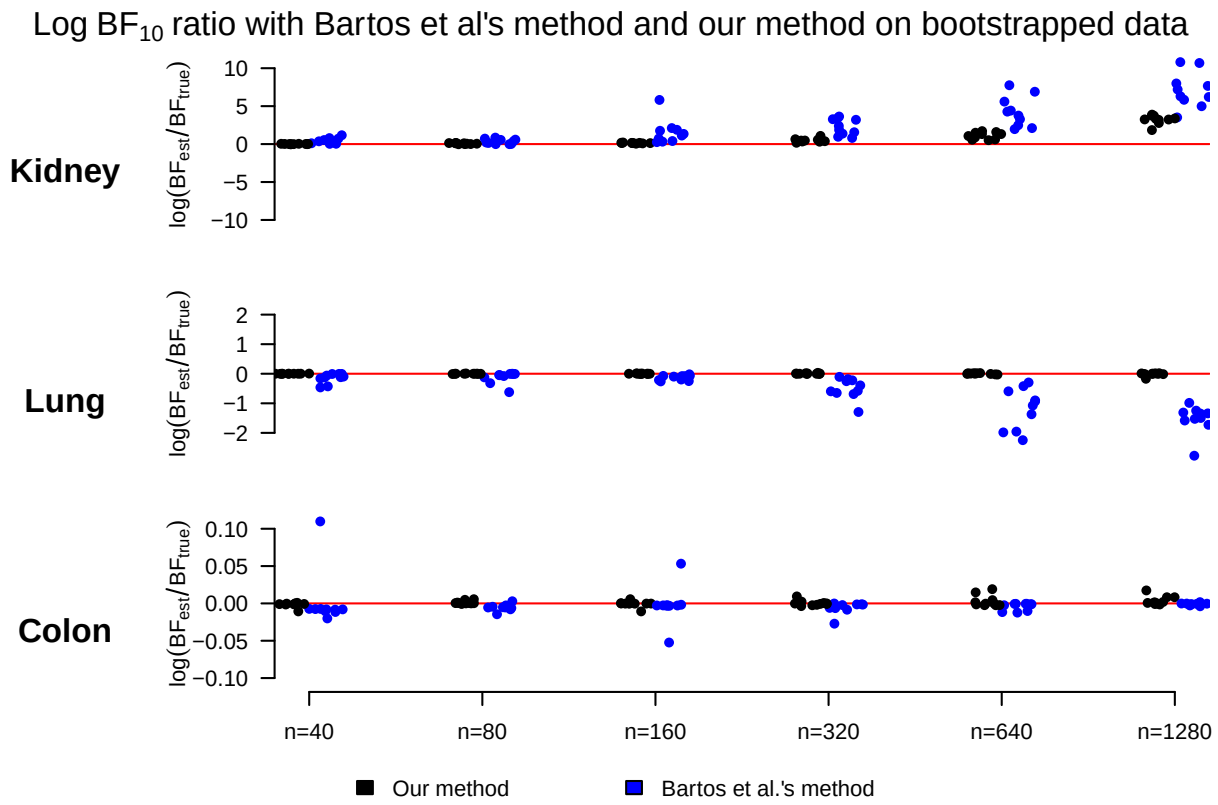

Figure E1

Log ratio of the estimated Bayes factor using our method (black) or the method by [4] (blue) and the true Bayes factor as a function of the sample size of the bootstrapped data sets. The red horizontal line shows the optimum.

resulting Bayes factors. For our method, we computed the normalized Efron approximation to the true log Cox partial likelihood and for the method by [4] we computed the normalized log Normal likelihood (i.e., a log Normal density) on a fine grid of β values for each bootstrapped data set. The normalization was done to make the log likelihoods of the two methods comparable. Subsequently, we computed the similarity of the two normalized log likelihood distributions by using the Kullback-Leibler divergence (KLD; [6]). The results are shown in Figure E2. It can be seen that in most cases, the KLD is very low, indicating that the distributions were very similar. As the sample size increased, the KLD decreased, indicating higher similarity of the normalized log likelihoods for larger samples.

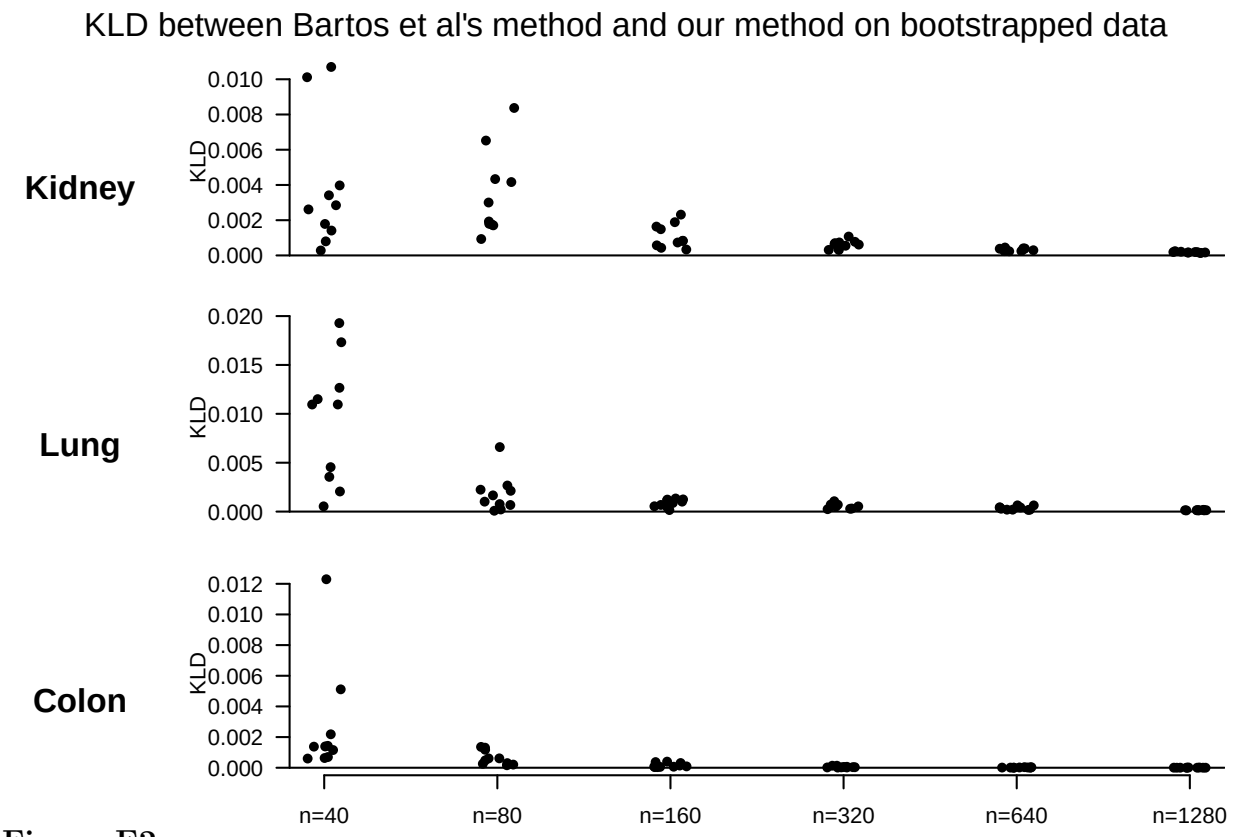

Figure E2

Kullback-Leibler divergence (KLD) between the normalized likelihood of our method and the normalized likelihood of the method by [4] as a function of the sample size of bootstrapped data sets.