STATISTICS FOR THE PEOPLE

INTERNATIONAL JOURNAL OF
RADIATION ONCOLOGY • BIOLOGY • PHYSICS
www.redjournal.org

Understanding the Differences Between Bayesian and Frequentist Statistics

Isabella Fornacon-Wood, MRes,* Hitesh Mistry, PhD,* Corinne Johnson-Hart, PhD,† Corinne Faivre-Finn, PhD,*,‡ James P.B. O'Connor, PhD,*,§ and Gareth J. Price, PhD*,†

*Division of Cancer Sciences, University of Manchester, Manchester, United Kingdom; †Departments of Medical Physics; ‡Clinical Oncology; and §Diagnostic Radiology, The Christie Hospital NHS Foundation Trust, Manchester, United Kingdom

Received Jul 9, 2021; Accepted for publication Dec 9, 2021

Case Vignette

Changes to radiation therapy workflows happen continuously as technology and techniques are optimized. One may wonder what the clinical outcomes of such changes are, or if there is any effect at all, because formal evaluation through a clinical trial rarely takes place. An example might be the modification of treatment protocols such as those used in image guided radiation therapy (IGRT). Suppose a large cancer center uses an IGRT action threshold during lung cancer patient setup. Patients whose position in their daily cone beam computed tomography (CBCT) is more than a threshold distance from that in their radiation therapy planning CT scan have their position corrected before treatment, whereas those with smaller offsets are treated without correction. The center updates its protocol to reduce the action threshold so that smaller setup errors are corrected. Published results suggest this change may have an impact on patient survival, and the team wants to determine whether this is the case in their center. Statisticians are tasked with identifying the appropriate statistical methodologies for such an evaluation as well as considering whether the approach could be embedded into routine practice to monitor the impact of future changes in clinical management.

Introduction

Just like Liverpool versus Manchester United, the Yankees versus the Red Sox, or Coke versus Pepsi, there are 2 main schools of statistics, and you may have heard the proponents of each noisily arguing their respective benefits. The first is the frequentist approach, which dominates the medical literature and consists of null hypothesis significance testing (think *P* values and confidence intervals). The other is the Bayesian approach, governed by Bayes' theorem. The fundamental difference between these 2 schools is their interpretation of uncertainty and probability[1]: the frequentist approach assigns probabilities to data, not to hypotheses, whereas the Bayesian approach assigns probabilities to hypotheses. Furthermore, Bayesian models incorporate prior knowledge into the analysis, updating hypotheses probabilities as more data become available. The goal of this article is to educate readers about the differences between frequentist and Bayesian inference, discuss potential advantages or disadvantages of each approach, and use the case vignette to highlight how these 2 methods may be implemented in a real-world example in a radiation therapy clinic.

Corresponding author: Gareth J. Price, PhD; E-mail: gareth.price@manchester.ac.uk

This research has been supported by Cancer Research UK via funding to the Cancer Research UK Manchester Center (C147/A18083 and C147/A25254), to RadNet Manchester (C1994/A28701), and to Professor O'Connor (C19221/A22746). Professor Faivre-Finn and Professor O'Connor are supported by the National Institute for Health Research (NIHR) Mancheser Biomedical Research Centre.

Disclosures: none.

Int J Radiation Oncol Biol Phys, Vol. 112, No. 5, pp. 1076–1082, 2022
0360-3016/$ - see front matter
https://doi.org/10.1016/j.ijrobp.2021.12.011

Introduction to Frequentist Statistics

Frequentist statistics is all about probability in the long run; the data set collected and analyzed is one of many hypothetical data sets addressing the same question, and uncertainty is due to sampling error alone. For example, the probability of getting heads when flipping a coin in the long run is 0.5; if we flip the coin many times, we would expect to see heads 50% of the time, whereas if we had flipped the coin only a few times we could reasonably expect to observe a different distribution (eg, all heads) just by chance.

Frequentist inference begins by assuming a null hypothesis to be true before data are collected (eg, that there is no effect of a particular treatment on survival). Investigators then collect data, analyze them, and ask, "How surprising is my result if there is actually no effect of the treatment on survival?" The data would be surprising if there was a low probability by chance alone of obtaining another data set at least as extreme (ie, far away from the null hypothesis and unlikely to occur by chance) than that collected (eg, showing a large difference in survival between patients having different treatments when our null hypothesis states there is no difference). If this is the case, then the collected data are considered unlikely under the null hypothesis and we can reject it, inferring that the null hypothesis does not adequately explain our data and that something else (eg, the new treatment) must account for our results.

This probability of obtaining another data set as extreme as the one collected is known as the *P* value. The *P* value is often criticized for being misunderstood and misused in the field of medicine.[2,3] For example, in contrast to popular belief, the *P* value is not a measure of how correct a hypothesis is, nor is it a measure of the size or importance of an effect.[3] In particular, large *P* values do not provide evidence of no effect.[4] A *P* value is simply the probability of obtaining another data set at least as extreme as the one collected by chance alone.

For example, suppose we run an analysis and get a highly significant *P* value of .001 for our treatment variable—how do we interpret this? Formally, there is a 0.1% chance of collecting data equal to or more extreme than this result if the null hypothesis were true—it would be surprising to collect these data if there is, in fact, no effect of treatment on survival. If we had run the analysis and got a *P* value of .46, then there is a 46% chance of collecting data equal to or more extreme than this if the null hypothesis is true—it would not be surprising to obtain these results if there is no effect of treatment on survival. This result is not evidence of no effect, however, because the inference began by assuming there would be no effect of treatment. The key here is that probabilistic statements (ie, *P* values) can only be made about the data, not about hypotheses or parameters (ie, the treatment effect).[5]

Reporting confidence intervals can improve the interpretation of results compared with a *P* value alone and can give information on the size and direction of an effect.[6] A 95% confidence interval tells us that if we were to repeat the experiment over and over (remember, frequentist statistics are long run), 95% of the computed confidence intervals would contain the true mean.[7] This is different than saying there is 95% chance the true mean lies within the interval, because frequentist statistics cannot assign probabilities to parameters—the true mean either lies within the interval or it does not.[8]

Introduction to Bayesian Statistics

Bayesian statistics are named after the Reverend Thomas Bayes, whose theorem describes a method to update probabilities based on data and past knowledge. In contrast to the frequentist approach, parameters and hypotheses are seen as probability distributions and the data as fixed. This idea is perhaps more intuitive because generally the data we collect are the only data set we have, so it does not necessarily make sense to perform statistical analysis assuming it is one of many potential data sets. Probability distributions summarize the current state of knowledge about a parameter or hypothesis and can be updated as more data becomes available using Bayes' theorem, presented in Equation 1.

$$p(\theta|Data) = \frac{p(Data|\theta) \cdot p(\theta)}{p(Data)} \quad (1)$$

The probability distribution that summarizes what is known about an outcome before a test or piece of information is obtained is known as the prior distribution, often just dubbed "the prior." Such an outcome may be the prevalence of disease or a specific diagnosis. The prior probability of the outcome θ (eg, of having the disease) is labeled $P(\theta)$ in Equation 1. The prior is one of the key differences between frequentist and Bayesian inference; frequentist analyses base their results only on the data they collect. The prior could be formulated by expert beliefs, historical data, or a combination of the two.

Consider now that one has a positive test for the disease. What is the probability of θ (having the disease), given these new data (ie, that the test was positive)? The notation for this scenario is $P(\theta|$ test +), where the "|" can be translated as "given" and we set the *Data* variable in Equation 1 to *test* +. In Bayesian terminology, this probability is known as the posterior distribution (or posttest distribution) and summarizes what is known about the outcome using both the prior information and the new data.

Among all potential patients, regardless of whether they have the disease, the probability of a positive test (ie, true positive plus false positive) is P(test+). Therefore, the relative probability of having a positive test if a patient is truly positive versus the probability of someone randomly selected from the population (who may or may not have the disease) having a positive test is the ratio $P(\text{test}+|\theta)$ / P(test +). The posterior probability of having the disease if you have a positive test is then the baseline prevalence of the

disease in the population (the prior probability) multiplied by this factor.

We can further illustrate how Bayes' theorem and the use of prior information can help to answer important questions using the example of COVID-19 testing. For this example, let's assume that a test for COVID-19 infection is guaranteed (100% chance) to detect the COVID-19 virus in someone who has the infection (P[test+| virus] = 1.0) and has a 99.9% chance of correctly identifying that someone does not have the virus (or a 0.1% chance of a false positive: P[test- | no virus] = 0.999 and P[test + | no virus] = 0.001).[9] That sounds like a high probability, but what we are really interested in is if you have a positive test, how likely is it that you actually have the virus (P[virus | test+]). We can calculate this probability using Bayes' theorem. As discussed, this probability depends on some prior information—how likely you were to have the virus (ie, P[virus]) before taking the test.

$$p(virus|test+) = \frac{p(test+|virus) \cdot p(virus)}{p(test+)} \quad (2)$$

If we assume the prior probability is the prevalence of the virus in the population at the height of the pandemic, 2%, then P(virus) = 0.02. Out of 1 million people, 20,000 will have the virus and 980,000 will not. If we test them all, 20,000 will have a true positive test, 979,020 will have a true negative test, and 980 will have a false positive test (980,000 × 0.001). Thus, there would be 20,980 positive tests in total (P[test+] = 20,980/1,000,000 = 0.02098), and the chance of having the virus given a positive test could be calculated using Equation 2: P(virus | test +) = (1.0 × 0.02)/0.02098 = 0.953, which can be approximated as 95%. In this setting, if the test is positive, one should believe the test.

However, what if the prevalence of COVID-19 was estimated to be much less, say, 0.2% (P[virus] = 0.002), such as during a period of a governmental stay-at-home order? In this scenario, when we test 1 million people, 2000 will have a true positive test, 997,002 will have a true negative test, and 998 will have a false positive test (998,000 × 0.001), giving 2998 positive test results, and P(test+) = 2988/1,000,000 = 0.002988. The probability of having the virus after a positive test now becomes P(virus | test +) = (1.0 × 0.002)/0.002988 = 0.669. In this scenario, one may truly question whether a positive test is diagnostic of infection, because the likelihood of a false positive is approximately 1 in 3.

Therefore, knowledge of the prior information (in this case, the prevalence of COVID-19 in a population) can alter the chance of having a virus when a test is positive from 95% to 67% when the sensitivity and specificity of the test remain unaltered. On the other hand, if we were testing hospitalized patients, for example, the prevalence would be expected to be much higher, and there would be a lower chance of obtaining a false positive result. This simple calculation highlights the importance of taking into account prior information, something a frequentist analysis does not do.

Sometimes formulating a prior is not that easy or clear. In such cases, the use of priors can be seen as a drawback to Bayesian inference, particularly when results depend on the chosen prior, and thus the analysis could be manipulated to get a positive result. In such cases, an "uninformative" prior that provides no additional information could be used, or multiple priors (eg, with either optimistic or skeptical assumptions) could be tested to determine the sensitivity of the results to particular priors. As with all analyses, it is vital that researchers are transparent in their methods and assumptions.

The posterior distribution captures our "updated" estimate of the probability of the outcome after incorporating new data, including our uncertainties, and can be analyzed to give various statistics, such as the mean and 95% credible interval. The 95% credible interval is different from a frequentist 95% confidence interval; it is the parameter range that has a 95% probability of including the true parameter value. The frequentist confidence interval is often misinterpreted in this way; however, one must remember that the 95% confidence interval assumes that the experiment is hypothetically repeated over and over and that 95% of the computed confidence intervals would contain the true mean. Perhaps more importantly, and one of the big advantages of Bayesian inference, is that the posterior distribution can also be used to directly calculate the probability of different hypotheses (eg, that one treatment is superior to another or that survival is improved by at least 3 months).

In this respect, Bayesian inference is more intuitive at its core and in closer alignment with our natural mode of probabilistic reasoning than frequentist inference. For example, we are more interested in the probability that 1 treatment is superior to another (Bayesian probability) than in the probability of obtaining certain data assuming the treatments are equal (frequentist null hypothesis). This advantage in interpretability remains even if our analysis uses an uninformative prior.

Case Study in Radiation Therapy

Let's look now at the real-world example in radiation therapy from our case vignette and compare the use of a frequentist and Bayesian approach to evaluating the clinical impact of a change in practice. In image guided radiation therapy (IGRT), an action threshold is often used as a decision threshold. At each daily fraction, the patient is set up on the radiation therapy treatment couch. They are then imaged using a CBCT system and their position is compared with the ideal position in the treatment plan. If the offset between the daily CBCT image and the planning CT image is greater than the action threshold, the couch is moved to better align the patient's position to that in the plan. If the offset is less than the threshold, the patient's setup is considered accurate enough and they receive their daily treatment without any shifts, the assumption being that setup errors less than the action threshold will not alter the clinical target

volume dose owing to the planning target volume margin or change organ-at-risk doses enough to have a clinical impact. Previous analyses of patients with lung cancer treated with IGRT have shown that the residual setup errors that remain after the action threshold has been applied (ie, positional errors that are less than the threshold and are considered acceptable enough to treat with) are associated with survival.[10] Patients with average residual setup errors that pushed the radiation therapy dose toward the heart were found to have worse survival than those who had average residual errors moving the dose away from the heart. In this previous study, the action threshold was 5 mm.

Let us assume that after this finding, the department decided to reduce the action threshold from 5 mm to 2 mm in the hope of ameliorating the effect. A year after implementing the change in protocol (postprotocol change), a physician wants to know the impact on clinical outcome. The hypothesis is that the effect of the average residual setup error direction (toward vs away from the heart) on patient survival will be decreased after reducing the action threshold from 5 mm to 2 mm (ie, the hazard ratio [HR] between the 2 directions will have decreased), because only trivial offsets would remain with the new policy.

We have 2 simulated data sets: 1 before protocol change, where residual setup errors range from −5 mm to 5 mm, as in the original study, and 1 after protocol change, with errors from −2 mm to 2 mm. The direction of the average residual setup error is calculated and dichotomized as either closer to or farther from the heart. We want to know if the difference in survival between patients with average residual setup errors toward versus away from the heart observed in the prechange patients (5-mm action threshold) is reduced with the new 2-mm protocol (ie, the HR of death between toward and away patients is reduced). We fit survival models to the pre- and postprotocol change data separately and look at how the HR changes, adjusting for the clinical variables accounted for in the cited study[10]: performance status, age, prescribed radiation therapy fractions, and tumor volume. The hypothesis is that the HR should be closer to 1 in the postprotocol change data, because a smaller action threshold should lead to smaller residual setup errors, and hence, less difference in the magnitude of heart irradiation and subsequent toxicity.

First, we take a frequentist approach. We fit a survival model to the preprotocol change data and postprotocol change data separately. The null hypothesis is that there is no effect of setup error on survival, and our model will indicate how surprising the data we collect are if this is true. Before protocol change, we have an HR of 1.26 (95% confidence interval, 1.05-1.48), and $P = .013$; that is, the risk of death is 26% (with a confidence interval of 5%-48%) higher in the group of patients with the average setup error direction toward the heart compared with those with the average setup error direction away from the heart. If there were actually no effect of setup error on survival, it would be surprising to see an HR this extreme (ie, this far from 1.0). After protocol change, we have an HR of 1.07 (95% confidence interval, 0.91-1.27), and $P = .41$. This P value is greater than the widely used 5% threshold, so seeing these results is not particularly unexpected if there were no difference in outcome associated with setup error. However, remember that this is not evidence of no effect, so we cannot conclude that the effect of residual setup error on survival is eliminated. Indeed, if we look at the 95% confidence interval, the HR of the postprotocol change could be as high as 1.27 (ie, a 27% increase in the risk of death). If we had a very tight confidence interval around the null (HR = 1), we would be more confident that there was no effect, but with such a wide confidence interval, the result is not informative.

Now we can analyze the data using Bayesian statistics. We again fit a survival model to the postprotocol change data, but this time, we specify a prior. To assess the sensitivity of the results to the prior, we choose 3 priors—an uninformative, a skeptical, and an enthusiastic one. The uninformative prior lets the postprotocol change data alone drive the inference, like in a frequentist setting. The skeptical prior (ie, skeptical that the change reduces the impact of the setup error) uses the data about the impact of the residual setup errors on survival from before the new protocol was changed (the preprotocol change patient cohort). The use of such historic data is a common scenario as it allows us to include existing observations (eg, from previous clinical trials) in our analysis and, for example, indicate how confident we are in the quality of the evidence (for observational data we might include greater uncertainty in the prior distribution). Where we are investigating a change in practice, it also allows us to be cautious in our assessment of the intervention by starting from the assumption that patients will experience the historically observed outcomes. The enthusiastic prior (ie, enthusiastic that there will no longer be an impact of setup error) assumes that there will be no survival difference between patients with residual setup errors toward versus away from the heart before the model sees the postprotocol change data. Enthusiastic priors are typically used for sensitivity analyses to evaluate how strongly the choice of the prior influences the analysis result. Given enough data, both the enthusiastic and skeptical priors would eventually evolve to the same posterior distribution (conversely, experiments with few data are unlikely to move the prior by a large amount).

Figure 1 presents the probability distributions for the HR before and after protocol change for the different priors, calculated via the Bayes theorem (Equation 1). We know the prior distribution ($P[\theta]$, or the belief about the distribution of the HR of death between patients with average setup errors toward and away from the heart) and obtain the distribution of the likelihood function ($P[\text{Data}|\theta]$, the probability of observing our survival data for a range of possible HR values) from our survival model. This information allows us to calculate the numerator of the Bayes equation, but calculating the denominator, $P(\text{Data})$, is often very difficult for nontrivial (ie, real-life) situations, meaning it is often not

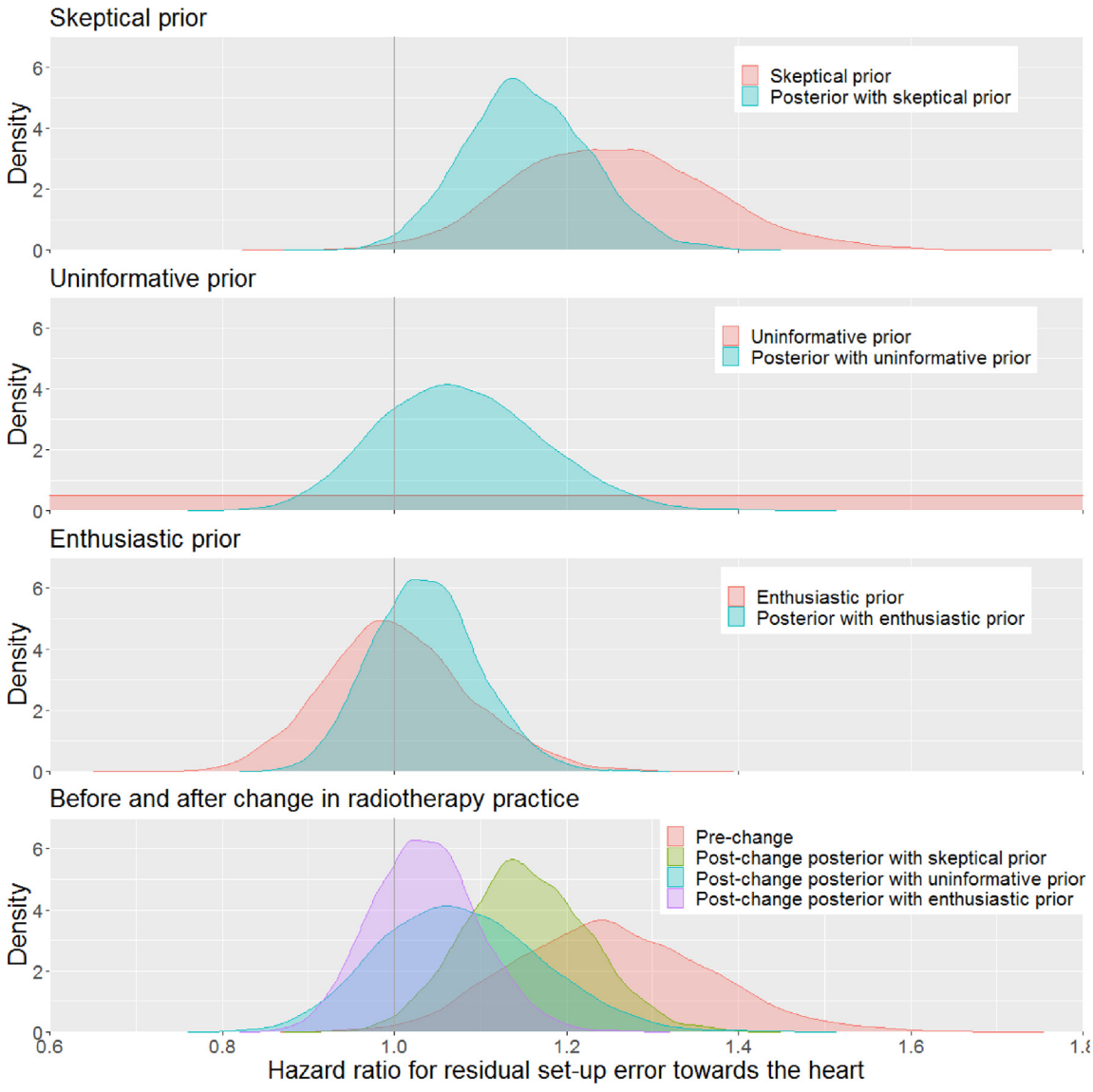

Fig. 1. Posterior distributions of the hazard ratio (HR) for residual setup error direction toward the heart. The red distribution in the top 3 plots shows the skeptical, uninformative, and enthusiastic priors used to calculate the HR distributions after protocol change, with the corresponding posteriors shown in blue. The bottom plot shows all 3 posterior HR distributions together with the HR distribution from before the protocol change (the prechange cohort).

possible to directly calculate the posterior distribution. Fortunately, we can instead use a computer to numerically sample a close approximation to it, typically using a technique called Markov chain Monte Carlo (MCMC) sampling. Briefly, MCMC iteratively samples different positions (ie, parameter values) in a hypothetical posterior distribution, using a set of rules to decide how to move from 1 position to the next (the rules are the Markov chain). At each new position we calculate the ratio of the new posterior probability to that in the previous position based on the obtained data. As Equation 3 shows, taking the ratio of the probability at the 2 positions means that we cancel out of the part of the Bayes equation that we have difficulty calculating, meaning we can easily calculate the ratio for each successive position. If the ratio is greater than 1 (ie, the new position is more probable given the data), we accept the move, and if it is less than 1 (ie, the previous position was more probable given the data), we calculate a random number and accept the move if the ratio is larger than the random number, rejecting it otherwise. By repeating this process many times and keeping a record of the different accepted sample positions (the record is known as the MCMC trace), we obtain an increasingly accurate estimation of the posterior distribution as the histogram of how often each parameter position "wins" relative to the competing value. Most commonly used analytical software languages contain MCMC sampling libraries to calculate posterior distributions using more sophisticated implementations of this general approach.[11,12]

$$\frac{p(\theta_{new}|Data)}{p(\theta_{previous}|Data)} = \frac{\frac{p(Data|\theta_{new})\cdot p(\theta_{new})}{p(Data)}}{\frac{p(Data|\theta_{previous})\cdot p(\theta_{previous})}{p(Data)}} = \frac{p(Data|\theta_{new})\cdot p(\theta_{new})}{p(Data|\theta_{previous})\cdot p(\theta_{previous})} \quad (3)$$

We can see that the postprotocol change HRs shown in Figure 1 are different depending on which prior was chosen; the posterior calculated with the skeptical prior has the largest HR and is closest to the preprotocol change HR distribution, the posterior with the enthusiastic prior has the HR distribution closest to the null (HR = 1), and the posterior using the uninformative prior has an HR somewhere in the middle. This result shows us how choice of prior influences inference in the Bayesian setting. With the skeptical prior, there is a postprotocol change HR = 1.15 (credible interval, 1.02-1.3). With the uninformative prior, it is slightly less (1.08 [0.91-1.27]), and with the enthusiastic prior, it is even less (1.04 [0.91-1.16]). Another key advantage of the Bayesian approach is that the HR point estimate and 95% credible intervals are not the only information that can be obtained from the posterior distributions; we can also directly quantify the evidence for multiple hypotheses using probabilities, examples of which are summarized in Table 1.

Interpretation of the results is different depending on the prior used. For the skeptical prior, there is a high probability that an effect of residual setup error direction toward the heart exists after protocol change, because we have a high probability (0.988) of an HR >1. The probability that the HR is reduced after protocol change is lower, at 0.760. With the uninformative prior, there is a moderate probability both that an HR >1 exists after protocol change and that the HR is reduced. With the enthusiastic prior, there is a high probability that the HR is reduced after protocol change and a lower probability (0.712) that an HR >1 exists.

So which prior and interpretation do we trust? That depends on our beliefs prior to the analysis and can be subjective. Presenting how sensitive results are to the choice of prior is therefore important so readers can fully understand how the prior affects the results and how this in turn affects the study's interpretation. For example, the results in Table 1 show that even when using an enthusiastic prior that assumed changing the IGRT protocol would eliminate the increased risk of death in patients whose setup errors move the heart toward the high dose region, there was a reasonably high probability that a residual effect remained. On the other hand, even when using a skeptical prior that assumed there would be no change in the HR after the protocol change, there was a reasonably high probability that the HR was reduced. As such, we can be confident that this finding (ie, the HR was decreased by the protocol change, but a reduced effect still remained) is a real effect, whereas the frequentist analysis does not allow us to make this inference and would be at risk of researchers reaching the opposite conclusion if only looking at the *P* value.

Use of an uninformative prior lets the data alone drive the inference and gives results similar to a frequentist analysis, but a Bayesian analysis allows one to directly calculate the probabilities for potential hypotheses. Furthermore, Bayesian posteriors allow us to calculate the probability of many different hypotheses (eg, that the HR is >1, that the HR is >1.1, that the effect is reduced by 20%). If we wanted to do this in the frequentist setting, we would need to separately assess different hypotheses and consider if multiplicity corrections were required. Similarly, the ease with which Bayesian analyses can accommodate prior information also means that it is straightforward to collect more data after an analysis has taken place if results are inconclusive.

Unlike the somewhat rigid orthodoxy that has developed around the interpretation of frequentist analyses (eg, *P*-value thresholds), Bayesian probabilities need to be placed into context to aid subsequent decision-making. The probability of different (competing) hypotheses can be directly cited as evidence and used to inform clinical decisions (in our example, if we consider that there is a strong probability that a residual setup error effect remains, we might then want to reduce the action threshold further). Formal frameworks for decision-making based on Bayesian probabilities have also been developed that are more akin to the hypothesis-testing approaches used in the frequentist setting. Bayes factors, for example, assess the ratio of the likelihood of 1 hypothesis over an alternative hypothesis given the observed data. The higher the Bayes factor, the more likely that 1 hypothesis (in the numerator) is correct. Standardized scales of this factor have been developed to decide which hypothesis is the most compelling. A more detailed explanation of Bayes factors is given by both Goodman[13] and Schonbrodt and Wagenmakers.[14]

Conclusions

Both the frequentist and Bayesian approaches are useful for data analysis as long as they are interpreted correctly.

Table 1 Bayesian probabilities for the skeptical, uninformative, and enthusiastic priors

	Skeptical prior	Uninformative prior	Enthusiastic prior
P(HR reduced)	0.760	0.882	0.961
P(HR >1)	0.988	0.786	0.712

Abbreviations: HR = hazard ratio; *P*(HR reduced) = probability that the HR for patients with average residual setup error direction toward the heart (compared with those with average residual setup error directions away from the heart) was reduced after the protocol change; *P*(HR>1) = probability that the HR for the average residual setup error direction toward the heart was greater than 1 after protocol change.

The strength of the Bayesian approach is the incorporation of prior information and the ability to directly calculate the probability of different hypotheses from the posterior distribution. It is more in line with our natural mode of reasoning; if we are sure in our belief of something (eg, we have a strong prior formed from evidence of a large effect from a phase 3 clinical trial), data will have to be highly convincing to alter this belief, whereas if we are unsure either way (uninformative prior), inference is driven more by the data than by the prior, and our beliefs are more easily overturned. In contrast, the frequentist approach is driven by the data only, so there is no issue of subjectivity owing to the prior. Although on one hand, this means one cannot manipulate priors to get a particular result in a frequentist analysis, the findings can be manipulated in other ways, such as *P*-value fishing, and the interpretation of the results is less intuitive and can easily lead to misinterpretation. However you choose to analyze your data, it is of utmost importance to be transparent in the methodology and to correctly interpret the results in a manner consistent with the underlying statistical approach and any limitations in how the data were collected (eg, in a prospective randomized trial or from a retrospective observational cohort). Proper inference, and thus the clinical impact of one's analysis, depends critically on these principles.

Dos and don'ts

- Do define the data available and the clinical question and then select the most appropriate statistical analysis.
- Do investigate the influence of your prior on the result when performing a Bayesian analysis.
- Do ensure your interpretation of either a frequentist or a Bayesian analysis is correct, with particular attention to *P* values, confidence intervals, and credible intervals.
- Do not assume a Bayesian analysis will solve the problem of insufficient or poor-quality (ie, incorrectly recorded) data; the most important part of an analysis is the quality of the data.

References

1. Willink R, White R. The. *Disentangling Classical and Bayesian Approaches to Uncertainty Analysis*. New Zealand: Measurement Standards Laboratory of New Zealand; 2011.
2. Colquhoun D. The reproducibility of research and the misinterpretation of p-values. *R Soc Open Sci* 2017;4 171085.
3. Goodman S. A dirty dozen: Twelve P-value misconceptions. *Semin Hematol* 2008;45:135–140.
4. Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ* 1995;311:485.
5. O'Hagan T. Dicing with the unknown. *Significance* 2004;1:132–133.
6. Ranstam J. Why the P-value culture is bad and confidence intervals a better alternative. *Osteoarthr Cartil* 2012;20:805–808.
7. Hoekstra R, Morey RD, Rouder JN, et al. Robust misinterpretation of confidence intervals. *Psychon Bull Rev* 2014;21:1157–1164.
8. Sim J, Reid N. Statistical inference by confidence intervals: Issues of interpretation and utilization. *Phys Ther* 1999;79:186–195.
9. Dinnes J, Deeks JJ, Adriano A, et al. Rapid, point-of-care antigen and molecular-based tests for diagnosis of SARS-CoV-2 infection. *Cochrane Database Syst Rev* 2020;8 CD013705.
10. Johnson-Hart CN, Price GJ, Faivre-Finn C, et al. Residual setup errors towards the heart after image guidance linked with poorer survival in lung cancer patients: Do we need stricter IGRT protocols? *Int J Radiat Oncol Biol Phys* 2018;102:434–442.
11. Bürkner PC. brms: An R package for Bayesian multilevel models using Stan. *J Stat Softw* 2017;80.
12. Salvatier J, Wiecki TV, Fonnesbeck C. Probabilistic programming in Python using PyMC3. *Peer J Comput Sci* 2016;2016:1–24.
13. Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. *Ann Intern Med* 1999;130:1019–1021.
14. Schonbrodt FD, Wagenmakers E. Bayes factor design analysis : Planning for compelling evidence. *Psychon Bull Rev* 2018;25:128–142.