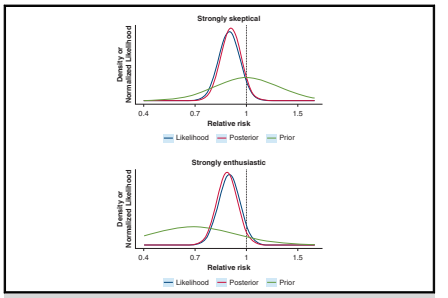
Bayesian post-hoc analysis of chemoradiation with and without surgery in patients with locally advanced squamous cell carcinoma of the esophagus

Nicolas Zhou, DO, MS,[a] Kyung Hyun Lee, PhD,[b] Claudia Pedroza, PhD,[b] and Wayne L. Hofstetter, MD[a]

Supplemental material is available online.

Randomized controlled trials (RCTs) are considered the greatest level of evidence in clinical research. In lower-prevalence cancers, where there is a paucity of RCTs, clinical practice may be heavily influenced by the results of a single randomized clinical trial. However, one must always carefully scrutinize the methodology of a trial before interpreting RCT results. In reality, RCTs are often more difficult to carry out and implement than studies in a controlled biological environment and might not replicate real-life decision-making strategies. Even though trial designs may be published beforehand, an RCT may not be conducted as originally designed if the design was unrealistic, does not reflect what clinicians and patients are willing to do, or if it encounters logistical obstacles that were not accounted for during the statistical design of the trial.[1] In addition, trials may be underpowered due to overestimation of treatment effect and unaccounted randomness inherent in real-world data or any unforeseen events in the implementation of the trial.

Most clinical trials are conducted and reported in a classical frequentist paradigm, which is with which clinicians are most familiar. The frequentist approach makes inferences about the intervention effect based solely on the current study's observed data without considering its plausibility or any previous studies. Frequentist results are regularly and inappropriately dichotomized into significant or not significant based on an established threshold by the investigators (ie, P value $< .05$). The Bayesian framework, in contrast, treats the true intervention effect as an unknown quantity for which an investigator can assign a prior belief about its most likely magnitude and plausible range based on known prior information. The investigator then updates their prior belief when encountering new evidence. We all have *prior beliefs* about whether a treatment is effective. When presented with new evidence, these beliefs are updated by incorporating these new data into what is known as the *posterior belief*, which is your updated belief now that you have incorporated the new data. These prior beliefs, new evidence, and posterior beliefs are expressed as *probabilities*. In the context of medical intervention, we believe that providing a probability of benefit and harm for a given intervention has the potential to facilitate personalized decision-making opportunities for patients, and these can only be obtained from a Bayesian analysis. From a Bayesian framework, the conversation around whether a medical treatment is effective is not yes or no but what is the probability of the treatment achieving desired outcome compared with control, and the probability of magnitude of effect.



Posterior probability of mortality at 5 years after trimodality with 5 different priors.

CENTRAL MESSAGE

Post-hoc Bayesian analysis is useful for evaluating the probability of treatment benefit in otherwise-rigid dichotomous interpretations of study results.

See Commentary on page 694.

From the [a]Department of Thoracic and Cardiovascular Surgery, The University of Texas MD Anderson Cancer Center; and [b]Center for Clinical Research and Evidence-Based Medicine, McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, Tex.

Generous philanthropic donations were received from the Mason Family Research Fund (N.Z).

Received for publication July 30, 2021; revisions received Oct 20, 2021; accepted for publication Nov 23, 2021; available ahead of print Nov 29, 2021.

Address for reprints: Nicolas Zhou, DO, MS, Department of Department of Thoracic and Cardiovascular Surgery, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Blvd, Houston, TX 77030 (E-mail: NZhou@MDAnderson.org).

J Thorac Cardiovasc Surg 2022;164:688-93
0022-5223/$36.00

https://doi.org/10.1016/j.jtcvs.2021.11.056

The most important questions for clinicians incorporating clinical trial results are often "Do these results help me make a decision with my patients and improve their outcome?" along with "How much confidence do I have in the results of the trial, and is there evidence that this trial applies to my population?" The reality is that the treatment benefit is rarely ever the same across different populations. The benefit of treatment should be more accurately understood as a spectrum of probabilities rather than a simple dichotomous outcome. To demonstrate a more nuanced understanding of the methodology in an RCT and how it affects our confidence in the results, we reanalyzed a highly cited trial by Stahl and colleagues[2,3] for esophageal squamous cell cancer under a Bayesian framework. We conducted a post-hoc Bayesian analysis to calculate the probability of any treatment benefit and of a clinically important treatment effect to demonstrate the differences between Frequentist and Bayesian interpretation of the results.

EVALUATION OF CLINICAL TRIAL METHODS

We will use a landmark study in esophageal squamous cell carcinoma (ESCC) as an example. The Stahl trial was a randomized trial that compared chemoradiation therapy followed by surgery (trimodality therapy [TMT]) with definitive chemoradiation therapy only (bimodality therapy [BMT]). The authors concluded in this trial that survival outcomes were equivalent regardless of whether patients were treated with surgery or without surgery. However, there are several flaws within the design of the trial from a frequentist perspective that can lead to this conclusion. We first summarize several methodologic flaws with the trial design that could affect the interpretation of benefit in a population outside of this study and then show how Bayesian analysis can provide more nuance to the results.

First, randomization occurred before the start of chemoradiation therapy. The decision to perform an esophagectomy after chemoradiation cannot be optimally determined before observing the effect of chemoradiation on the patient's fitness for surgery. This difficulty in implementation was reflected in the intent-to-treat analysis, as only 57 of 86 patients (66%) in the TMT arm actually underwent resection. Similarly, in the BMT cohort, where the intent was to avoid surgery, 5 of 86 patients (6%) underwent surgery resulting in substantial contamination of effect. This unintended cross-over increases the likelihood of an equivalent outcome between the 2 arms.

Second, the expected equivalence limit used in this study was a delta of 15% at 2 years. We argue that this parameter is too wide, as echoed by Ruhstaller and colleagues.[4] The sample size calculation is based on this subjective equivalence limit, and an overestimation of treatment effect lowers the sample size needed to detect a difference, or in this case, equivalence, but that makes the clinical relevance of this study subjective.

Third, even with the aforementioned parameters, this may have been underpowered to be an equivalence trial. We calculated a sample size based on the stated trial assumptions, assuming a 20% survival rate, a 15% equivalence margin, 80% power, and a significance level of .05. We estimated that a total of 244 patients would be needed, more than the 86 patients in each arm in this trial and the 100 patients per group in the power calculation before the adaptive design stated by the authors. While the authors used an adaptive design to stop the trial early, the actual adaptive design was not disclosed and hence cannot be evaluated for potential biases.[5]

Fourth, the trial reported an 11% perioperative mortality rate in TMT, which is greater than the generally accepted 3% to 5% that is expected for this procedure.[6] Given these concerns, the results of this study likely underestimated the benefit of TMT in patients with locally advanced ESCC. Thus, we used a post-hoc Bayesian analysis method to evaluate the probability of treatment benefit of TMT, as well as the probability of relative and absolute risk reduction in mortality at various follow-up years.

BAYESIAN STATISTICAL ANALYSIS

Bayesian analyses update the evidence of treatment benefit or harm. We start with a prior probability of treatment effect of TMT or BMT represented by relative risk (RR) for mortality with BMT as a baseline reference. This information could be derived from previous clinical trials, clinical knowledge, or a neutral prior to indicate a priori equal likelihood of either group being superior. The results from the trial by Stahl and colleagues can then be combined with each of the prior probabilities to create the posterior probabilities. These probabilities allow us to report point estimates and a (95%) credible interval (CrI) of treatment effect and calculate the probabilities of a treatment effect of a specific magnitude (ie, probability of relative risk reduction [RRR] and absolute risk reduction [ARR]).

The probabilities of RRR and ARR at different time points were calculated. We want to know the probability of RRR and the ARR because these become important discussion points with the patient when deciding whether the morbidity of an esophagectomy is worth the reduced risk of death from cancer. We used 5 different priors to explore how the study results vary depending on doubt or confidence the patients and clinicians might have regarding treatment benefit of a given treatment arm. We extracted the number of deceased patients in each group at 2, 3, 5, and 10 years after enrollment.[2,3] For each outcome, separate models were run for each prior distribution. We used a normal approximation for the observed log RR. All prior distributions were also assumed to follow a normal distribution in the log RR scale. (This assumption results in a normal posterior as well.) The likelihood was solely based on the observed mortality data (Table E1). For validation purposes and calculation of ARR, we fitted the models using Markov chain Monte Carlo methods described by Goligher and colleagues.[7] We calculated the probabilities of treatment benefit from TMT, defined as the probability that TMT provides any benefit (RRR >0%). To evaluate larger treatment benefits, we also calculated the probabilities of RRR >10% and >20%. In addition to RRR, we also evaluated the probability of ARR under each prior distribution with a fixed baseline mortality risk of 65% (BMT mortality at 2 years). The probabilities of ARR of 2%, 4%, and 6% were estimated under each prior distribution. The posterior probabilities are summarized with a mean estimate and a 95% CrI. Statistical analyses were performed using the metaphor and R2jags packages[8,9] in R software, version 3.6.2.[10]

DERIVATION OF PRIORS

Choosing values for the prior probabilities can be controversial[11] because they represent different prior beliefs that can lead to different conclusions. Therefore, we devised 5

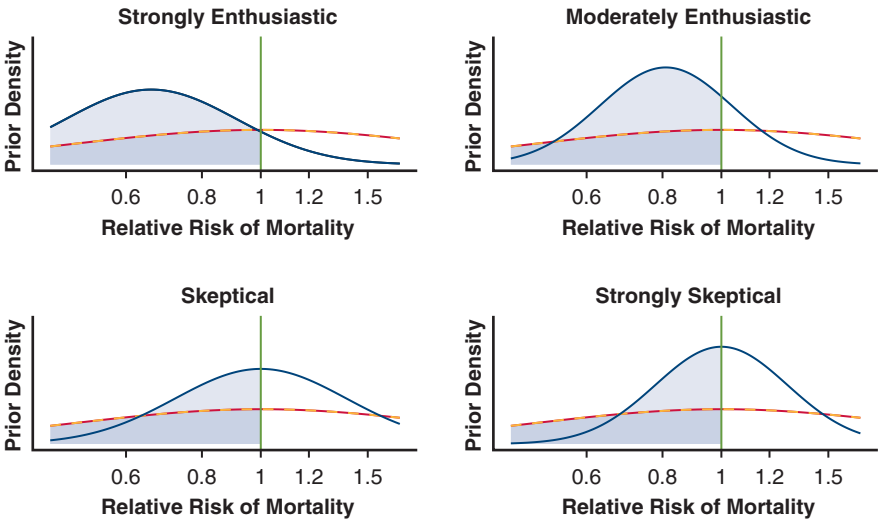


FIGURE 1. Five different prior probabilities for relative risk of mortality for TMT with reference to BMT: strongly enthusiastic, moderately enthusiastic, skeptical, strongly skeptical, and neutral prior. Neutral prior is indicated by *orange and black dotted lines* in every graph. The *shaded area* has a relative risk of mortality ≤1, indicating treatment benefit of TMT. *TMT*, Trimodality therapy; *BMT*, bimodality therapy.

different representations of prior information[12]: neutral (minimally informative), strongly enthusiastic (of TMT benefit), moderately enthusiastic, skeptical (no treatment benefit from TMT), and strongly skeptical priors (Figure 1). These priors act as sensitivity analyses that allow us to determine how the study results vary depending on the prior confidence of the clinicians in the benefits of a given treatment plan. In other words, we first start with a belief about the effectiveness of the treatment (prior probability). Then via a clinical trial, we obtain the data to estimate the probability of observing the outcome of interest (likelihood). This new information updates our initial belief about the effectiveness of the treatment one way or the other; hence, producing posterior probability. The reason we have multiple priors is to represent the real-world enthusiasm or skepticism that clinicians may have about a certain treatment before a clinical trial is done. The center of the distribution represents the most likely size of the treatment effect (ie, RR), and the width of the distribution represents the certainty around this value (eg, tighter curves represent more certainty whereas wider ones represent more uncertainty). The area under the curve to the left of any value (Figures 1 and 2) represents the probability of the treatment effect being less than that value (eg, the probability of RR < 1 is the area to the left of 1, which is identical to the probability of RRR >0%). Although we expect TMT to be beneficial, we first conservatively used a neutral prior to express the lack of evidence favoring TMT. The RR was centered at 1.0 (0 in log RR scale) and a 95% CrI of 0.25 to 4.0, encompassing the largest likely effect size for major outcomes in randomized trials.

We specified 2 different enthusiastic priors: strongly enthusiastic and moderately enthusiastic, with strongly enthusiastic having a lower RR than moderately enthusiastic. For the strongly enthusiastic prior, RR was centered at 0.66, which was derived from the clinical trial by Bosset and collegues.[13] This was the largest trial completed before the trial by Stahl and colleagues. For the moderately enthusiastic prior, we used Stahl colleagues' hypothesized 2-year mortality rate of 65% for TMT with BMT as a reference.[2,14] The RR for moderately enthusiastic prior was 0.81. The standard deviations (SDs) of each prior were 0.25 and 0.13, respectively. With SDs of 0.25 and 0.13, we assumed an a priori 90% and 80% likelihood of RR < 1.0 for a strongly enthusiastic and moderately enthusiastic prior, respectively.

Likewise, we also created 2 different skeptical priors: skeptical and strongly skeptical, both centered at an RR of 1.0. We adopted the same SDs, 0.25 and 0.13, from enthusiastic priors for skeptical priors. The smaller SD of 0.13 was used for the strongly skeptical prior because a smaller SD denotes stronger skepticism.[15] Note that even though the RR for both skeptical priors was centered at 1, there was still a 50% chance that BMT increases the mortality rate. While some skeptics may consider surgery to be detrimental to the mortality of outcome, this has not been reported in literature, and thus RR of 1 was meant to convey that each arms has a 50% chance of increasing mortality rate.

The trial enrolled 86 patients in each arm. The mortality rates in each arm at years 2, 3, 5, and 10, as well as the observed RRs, are summarized in Table E1. On the basis of each of the 5 priors, the 5 prior probabilities of RR < 1, <0.9, <0.8, and <0.67 (RRR >0%, >10%, >20%, >33%) for TMT are summarized in Table 1.

POSTERIOR PROBABILITY ANALYSES

The posterior probabilities of RR in mortality at 2 years under each prior were calculated (Table 2). Under a minimally informative (neutral) prior, the posterior estimate of RR at 2 years was 0.93 (95% CrI, 0.74-1.17), indicating an estimated 7% reduction in mortality in TMT compared

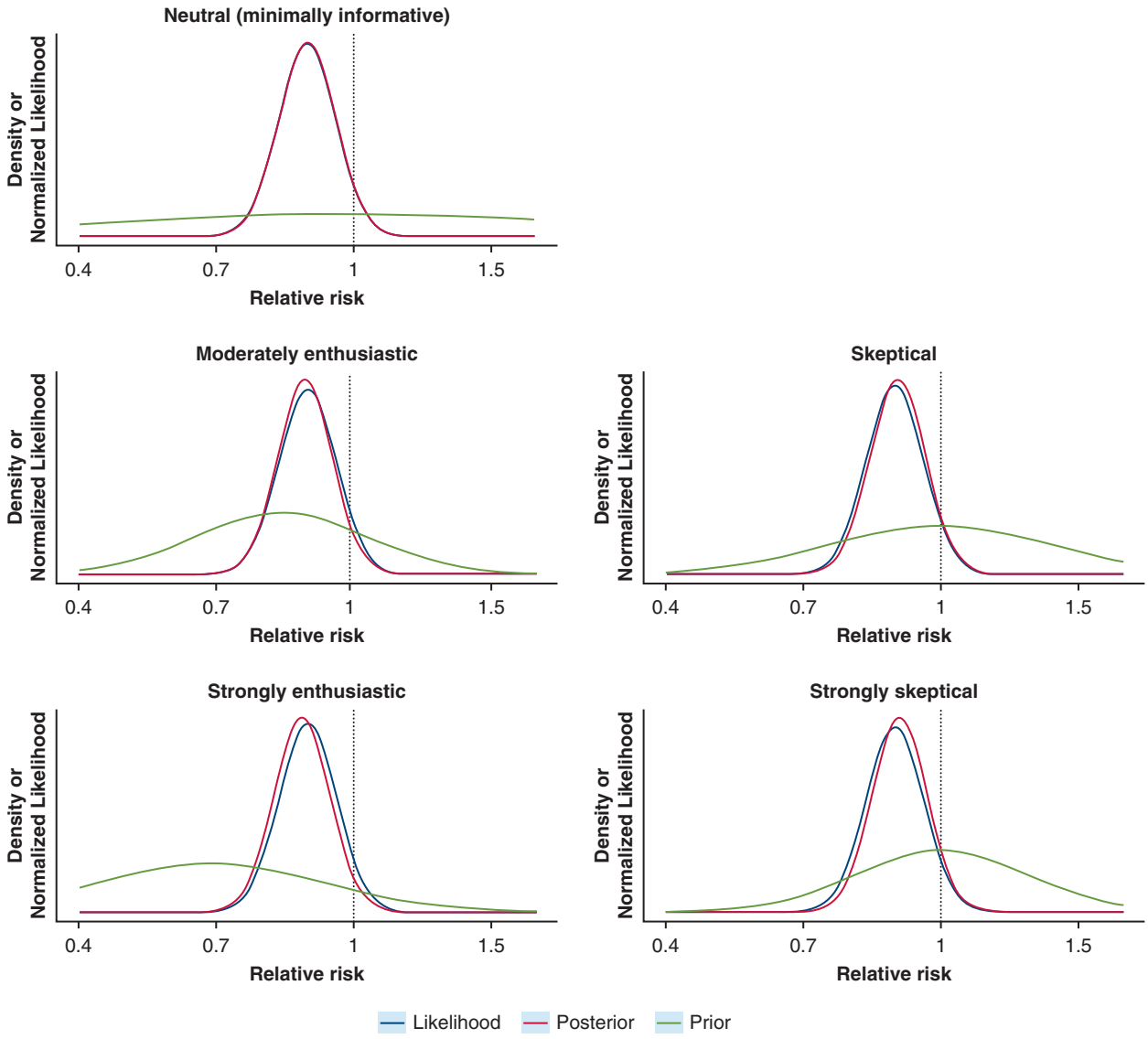


FIGURE 2. Five-year follow up posterior probabilities with 5 different priors; the prior belief (*blue*) is combined with the observed outcome likelihood (*red*) to generate the posterior probability (*green*) of relative risk of TMT compared with BMT. *TMT*, Trimodality therapy; *BMT*, bimodality therapy.

to BMT. For strongly and moderately enthusiastic priors, the posterior estimates of RR were 0.89 and 0.91, indicating an 11% and 9% reduction in mortality for TMT, respectively. For skeptical and strongly skeptical priors, the posterior estimate was both 0.94. The greatest probability of treatment benefit with TMT was seen at 5 years. With the strongly enthusiastic prior, the posterior probability of RRR >10% was 72.2%. Even with a strongly skeptical prior, the 5-year posterior probability of any treatment benefit for TMT was 93.8%. Figure 2 graphically

TABLE 1. Prior probability of RR and RR reduction in mortality with 5 different priors for TMT in comparison with BMT

Prior belief	Mean RR	SD of log (RR)	Prior probability of treatment benefit (in mortality) greater than or equal to specified threshold (%)			
			RR < 1	RR < 0.9	RR < 0.8	RR < 0.67
Neutral	1	0.7	50	44	37	28
Strongly enthusiastic	0.66	0.32	90	83	72	51
Moderately enthusiastic	0.81	0.25	80	66	48	22
Skeptical	1	0.32	50	37	25	11
Strongly skeptical	1	0.25	50	34	19	5

RR, Relative risk; *SD*, standard deviation.

TABLE 2. Posterior probability of treatment benefit (in mortality) greater than or equal to specified threshold based on 5 different priors at follow-up time of 2, 3, 5, and 10 years

	2 y			3 y			5 y			10 y		
Prior belief	Posterior median RR (95% CrI)	RRR >0%	RRR >10%	Posterior median RR (95% CrI)	RRR >0%	RRR >10%	Posterior median RR (95% CrI)	RRR >0%	RRR >10%	Posterior median RR (95% CrI)	RRR >0%	RRR >10%
Neutral	0.93 (0.74-1.17)	73.3	38.7	0.91 (0.76-1.09)	84.3	45.7	0.87 (0.74-1.03)	94.7	63.4	0.92 (0.81-1.05)	89.0	36.8
Strongly enthusiastic	0.89 (0.72-1.11)	84.8	53.0	0.88 (0.74-1.06)	90.9	57.3	0.86 (0.73-1.01)	97.1	72.2	0.91 (0.80-1.03)	92.9	45.1
Moderately enthusiastic	0.91 (0.74-1.12)	82.3	47.5	0.89 (0.75-1.07)	89.4	52.6	0.87 (0.74-1.01)	96.5	68.3	0.91 (0.80-1.04)	92.1	41.9
Skeptical	0.94 (0.75-1.16)	72.3	35.9	0.91 (0.76-1.09)	83.5	43.0	0.88 (0.75-1.03)	94.2	60.6	0.92 (0.81-1.05)	88.6	35.0
Strongly skeptical	0.94 (0.76-1.16)	71.6	33.6	0.92 (0.77-1.09)	82.9	40.8	0.89 (0.76-1.03)	93.8	58.3	0.93 (0.81-1.05)	88.3	33.6

Posterior probability of treatment benefit (in mortality) greater than or equal specified threshold based on five different priors at follow-up time of 2, 3, 5, and 10 years. *RR*, Relative risk; *RRR*, relative risk reduction; *CrI*, credible interval.

demonstrates the prior distribution, likelihood, and posterior probability distribution in a RR scale for the 5 different priors and posttreatment at 5 years.

A Markov chain Monte Carlo simulation was performed to calculate the ARR with a fixed baseline mortality rate of 65%, which was the mortality rate of BMT at 2 years. Table E2 describes the probability of at least a 2%, 4%, and 6% ARR for each prior at 2, 3, 5, and 10 years after enrollment. At 2 years after enrollment, the probability of an ARR > 2% was 66% with neutral prior, 75% with moderately enthusiastic prior, and 63% with strongly skeptical prior. At 5 years after enrollment, the probability of an ARR >2% was at least 88% regardless of priors, and the probability of an ARR >6% was at least 72% with enthusiastic priors. Note that, at 10 years, the probabilities of mortality reduction were lower than that after 5 years: the probability of an ARR >4% was 62% under neutral prior, at least 67% under enthusiastic priors and at least 60%

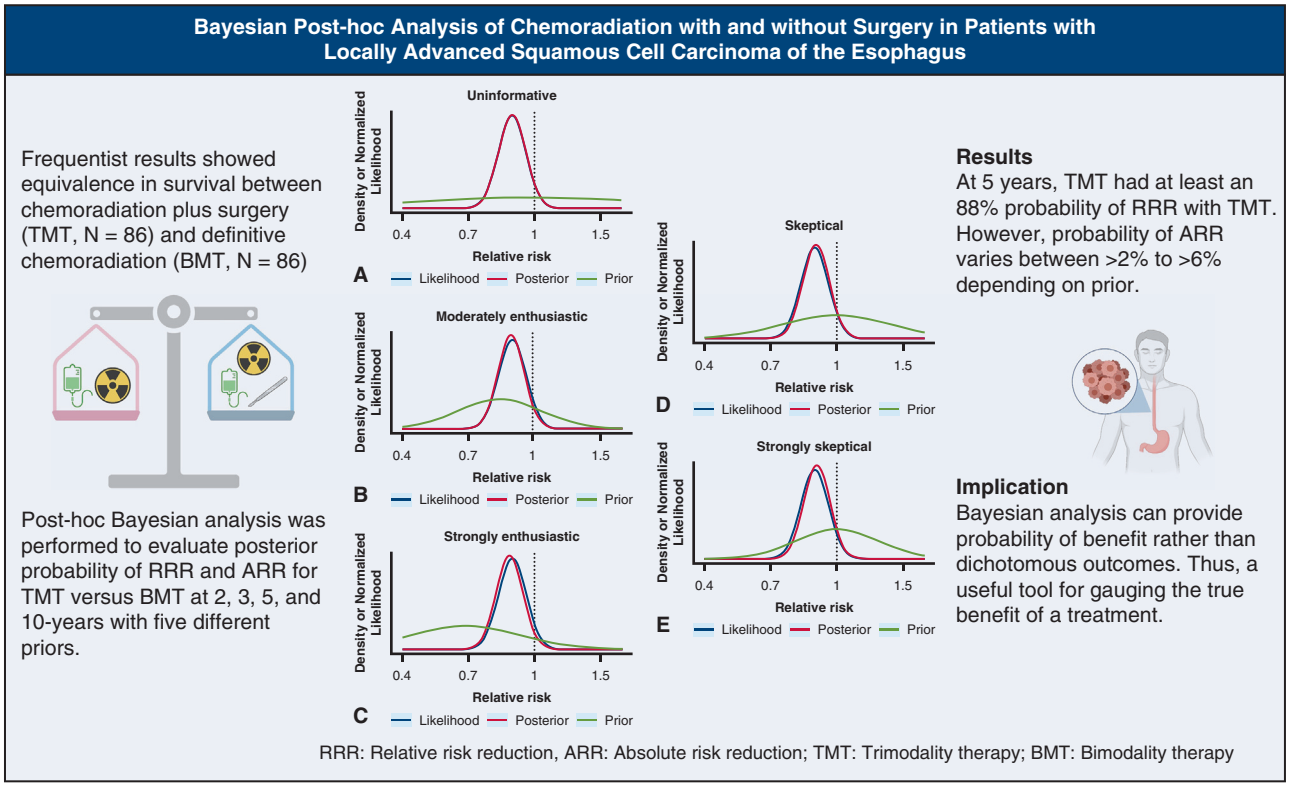


FIGURE 3. Visual abstract depicting the main message of the study. *TMT*, Trimodality therapy; *BMT*, bimodality therapy; *RRR*, relative risk reduction; *ARR*, absolute risk reduction.

under skeptical priors. We have provided the codes used to generate these analyses in R (Online Data Supplements 1-3).

Limitations

Bayesian methods are not without limitations because the use of priors is subjective, even when derived from other high-evidence studies. In our study, we attempted to minimize this limitation by using multiple different priors to evaluate the array of probabilities with a given prior, including the parameters as indicated by Stahl and colleagues. Another limitation is that this post-hoc Bayesian analysis could not account for the large contamination of effects due to the randomization schema or the higher-than-normal perioperative mortality rate.

DISCUSSION

Our Bayesian post-hoc analysis methods provide a probabilistic view of the trial results from Stahl and colleagues (Figure 3). Rather than a simple yes-or-no answer to the treatment benefit of TMT based on the investigators' hypothesis testing, we provided the probability and magnitude of the treatment benefit at 2, 3, 5, and 10 years after enrollment based on both known prior information and a neutral stance on treatment benefit. Our results can help patients with ESCC quantify the probability of survival benefit with TMT compared with BMT. In addition, because esophagectomy is a major surgery with significant risk for morbidity and mortality, patients and their caregivers can use the probability of ARR to decide whether the risk reduction in survival is worth the risk of surgery. While these probabilities are likely different in the contemporary era, Bayesian analysis is a useful tool in helping surgeons and patients assign numeric values in estimating risk trade-offs.

While we agree with the conclusion of the interpretation of the trial outcomes based on the frequentist design, we feel that issues with the trial design lead to an overly simplified answer that did not adequately convey the whole result of the trial. While there are now newer and larger trials on the horizon, the use of statistics continues to be mostly frequentist. This is in large part due to clinician's lack of familiarity with interpreting and using Bayesian statistics, but we believe that Bayesian analysis conveys additional information to trials and is a useful tool in patient care. Bayesian and frequentist statistical paradigms report information in different ways, but both are helpful for clinicians and patients in a personalized decision-making process based on each patient's risk tolerance. From patients' perspective, the decision for prolonged life versus quality of life is a complex individualized decision between treatment toxicity, morbidity, possible risks and benefits, and many others. These decisions and values are personal and situation-dependent.[16] To be able to produce a probability of the treatment benefit and magnitude of benefit for a specific outcome, such as survival, can be invaluable in complex decision-making discussions.

Conflict of Interest Statement

The authors reported no conflicts of interest.

The *Journal* policy requires editors and reviewers to disclose conflicts of interest and to decline handling or reviewing manuscripts for which they may have a conflict of interest. The editors and reviewers of this article have no conflicts of interest.

References

1. Thall PF. *Statistical Remedies for Medical Researchers*. Cham: Springer Nature Switzerland AG; 2020.
2. Stahl M, Stuschke M, Lehmann N, Meyer HJ, Walz MK, Seeber S, et al. Chemoradiation with and without surgery in patients with locally advanced squamous cell carcinoma of the esophagus. *J Clin Oncol*. 2005;23:2310-7.
3. Stahl M, Wilke H, Lehmann N, Stuschke M. Long-term results of a phase III study investigating chemoradiation with and without surgery in locally advanced squamous cell carcinoma (La-Scc) of the esophagus. *J Clin Oncol*. 2008;26:4530.
4. Ruhstaller T, Mueller A, Koeberle D, Thuerlimann B. Is there really enough evidence to abandon surgery after radiation chemotherapy in esophageal cancer? *J Clin Oncol*. 2005;23:8547-8.
5. Pallmann P, Bedding AW, Choodari-Oskooei B, Dimairo M, Flight L, Hampson LV, et al. Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Med*. 2018;16:29.
6. Raymond DP, Seder CW, Wright CD, Magee MJ, Kosinski AS, Cassivi SD, et al. Predictors of major morbidity or mortality after resection for esophageal cancer: a Society of Thoracic Surgeons General Thoracic Surgery database risk adjustment model. *Ann Thorac Surg*. 2016;102:207-14.
7. Goligher EC, Tomlinson G, Hajage D, Wijeysundera DN, Fan E, Jüni P, et al. Extracorporeal membrane oxygenation for severe acute respiratory distress syndrome and posterior probability of mortality benefit in a post hoc Bayesian analysis of a randomized clinical trial. *JAMA*. 2018;320:2251-9.
8. Viechtbauer W, Viechtbauer MW. Package 'Metafor'. The comprehensive R archive network. Package 'metafor'; 2015. Accessed July 20, 2020. http://cran.r-project.org/web/packages/metafor/metafor.pdf
9. Su Y-S, Yajima M, Su MY-S. System requirements J. Package 'R2jags'. R package version 0.03-08; 2015. Accessed July 20, 2020. http://CRAN.R-project.org/package= R2jags
10. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2013.
11. Spiegelhalter DJ. Bayesian methods for cluster randomized trials with continuous responses. *Stat Med*. 2001;20:435-52.
12. Quintana M, Viele K, Lewis RJ. Bayesian analysis: using prior information to interpret the results of clinical trials. *JAMA*. 2017;318:1605-6.
13. Bosset J-F, Gignoux M, Triboulet J-P, Tiret E, Mantion G, Elias D, et al. Chemoradiotherapy followed by surgery compared with surgery alone in squamous-cell cancer of the esophagus. *N Eng J Med*. 1997;337:161-7.
14. Stahl M, Wilke H, Fink U, Stuschke M, Walz MK, Siewert JR, et al. Combined preoperative chemotherapy and radiotherapy in patients with locally advanced esophageal cancer. Interim analysis of a phase II trial. *J Clin Oncol*. 1996;14:829-37.
15. Pedroza C, Han W, Thanh Truong VT, Green C, Tyson JE. Performance of informative priors skeptical of large treatment effects in clinical trials: a simulation study. *Stat Methods Med Res*. 2018;27:79-96.
16. Blinman P, King M, Norman R, Viney R, Stockler MR. Preferences for cancer treatments: an overview of methods and applications in oncology. *Ann Oncol*. 2012;23:1104-10.

Key Words: esophageal, squamous cell carcinoma, trimodality, bimodality, Bayesian, surgery, chemotherapy, radiation, chemoradiation

TABLE E1. Observed mortality data, n (%), from the trial by Stahl and colleagues

Mortality	TMT (n = 86)	BMT (n = 86)	RR
2 y	52 (60.5)	56 (65.1)	0.93
3 y	59 (68.6)	65 (75.6)	0.91
5 y	62 (72.1)	71 (82.6)	0.87
10 y	69 (80.2)	75 (87.2)	0.92

TMT, Trimodality therapy; *BMT*, bimodality therapy; *RR*, relative risk.

TABLE E2. Posterior probability that the true ARR was >2%, 4%, and 6% for each prior, assuming a baseline mortality rate of 65% at 2, 3, 5, and 10 years after surgery

	2 y			3 y			5 y			10 y		
	>2%	>4%	>6%	>2%	>4%	>6%	>2%	>4%	>6%	>2%	>4%	>6%
Neutral	65	54	43	77	65	51	90	81	68	79	62	42
Strongly enthusiastic	79	69	58	85	75	62	94	87	76	85	70	51
Moderately enthusiastic	75	65	52	82	71	57	93	85	72	83	67	48
Skeptical	64	53	41	75	62	48	89	79	65	77	60	40
Strongly skeptical	63	51	39	74	61	46	88	78	63	77	60	39