# How Similar is "Similar"?
## A Deeper Dive into Bucher Versus Bayesian Network Meta-Analysis

Kyle Fahrbach, PhD
Principal Statistician, Meta Research, Evidera

### Introduction

The companion article, *Bucher Versus Bayesian NMA Approaches for Indirect Treatment Comparisons: What Do HTA Agencies Want?*, in this issue of *The Evidence Forum* describes Bucher and Bayesian network meta-analysis (NMA) methods and how they are viewed by payers and health technology assessment (HTA) bodies. Here we elaborate on one of its key conclusions – that the two approaches usually yield similar, but not necessarily identical, findings. This potential for observed differences – even slight ones – can cause confusion and pose challenges around interpretation and use of the results. With this in mind, we examine what sort of numeric differences might be expected between the two methods and the possible causes.

As summarized in Table 1, there are three primary reasons why Bucher and Bayesian results might differ. Each reason is independent of the other two, and discrepancies between analyses can come from more than one source.

This potential for discrepancies between analyses is concerning on its own for a pharmaceutical company, however, it may be compounded by another challenge – that is, differences between the results in the company's (single) key trial and the results from a Bayesian NMA that includes that trial. This complication raises two additional points for consideration:

1. Why do Bayesian NMAs sometimes give results for a treatment comparison that differ from those reported in a single trial involving that comparison?
2. Why would a Bayesian NMA **give results that do not show statistical significance, unlike those from the head-to-head, single-trial comparison?**

We discuss all these issues in more detail below, using an invented treatment network for illustration.

Table 1.

| **Reason for Discrepancy** (Bucher vs. Bayesian) | **Potential Discrepancy** (Central Point Estimate) | **Potential Discrepancy** (95% Interval) | **Scenario** |
|---|---|---|---|
| Bayesian "noise" | Extremely Small | Extremely Small | Any |
| (Slight) differences in statistical modeling | Extremely Small to Small | Extremely Small to Small | Any |
| Difference in the estimates of heterogeneity between the analyses | Extremely Small to **Moderate** | Small to **Large** | Random-effects analysis only |


Kyle Fahrbach

## The Treatment Network

Figure 1 shows our invented network – one in which most studies have involved a common comparator (in this case, placebo), with one or two studies in the periphery. In this network, there is only *one study* per individual treatment comparison. For purposes of instruction, the network has no "closed loops", i.e., no instances where for any given comparison there is both direct and indirect evidence, or multiple paths of indirect evidence. (As noted in the companion article, networks with many "closed loops" are generally best analyzed with full network approaches rather than multiple, parallel Bucher analyses, although the latter remains an option.)

We use this network to provide examples of analyses using the Bucher and Bayesian approaches and to describe how discrepancies might arise when the techniques are applied side-by-side. As might be expected, the size of the potential discrepancies in estimates between the two methods is proportional to the complexity underlying that discrepancy. We begin, then, with the issue of "Bayesian noise."[a]

**Table 2a.** Estimates of Sucrosa vs. Pacifex
*(Mean Differences with 95% Confidence/Credible Intervals)*

| Analysis Technique | Network 1A |
|---|---|
| Bucher | 8.50 [4.63, **12.37**] |
| Bayesian (fixed-effects [FE] model) | 8.50 [4.63, **12.36**] |
| Bayesian (FE, increased # of simulations) | 8.50 [4.63, **12.37**] |

## Example 1. Discrepancies from "Bayesian Noise" for Mean Differences

In the simplest case – a two-study network (Network 1A) with an outcome such as a hazard ratio (HR) or mean difference – the similarity between results from a Bucher and a Bayesian (fixed-effects model) approach is obvious, but with a small catch. (Table 2a)

Specifically, both approaches give substantively identical point estimates and 95% intervals (as seen in the first two rows of Table 2a) – but they are not **completely** identical, as there is a 0.01 difference in the upper end of the 95% intervals. Similar discrepancies, on the order of a rounding error, often occur when conducting Bayesian NMAs. This is due to the analytical approach used in Bayesian estimation – the Markov chain Monte Carlo (MCMC) method – in which statistical models are used to simulate, and thereby predict outcomes of, treatment comparisons. This use of simulations means the Bayesian approach does not calculate estimates exactly, and changes to the key model inputs and/or the number of simulations can result in minor variations in the results.

The solution to this problem (if, indeed, one is deemed necessary) is usually simple: increase the number of simulations per chain (i.e., run 100,000 simulations instead of 50,000) and/or increase the number of MCMC chains (as each chain has its own unique set of simulations); as the total number of simulations increases, random noise will decrease.

[a] This section is dedicated to every pharmaceutical company who has asked why a result changed by 0.01 after an update.

**Figure 1.** Single-Study-Per-Comparison Network (Mean Difference Example)

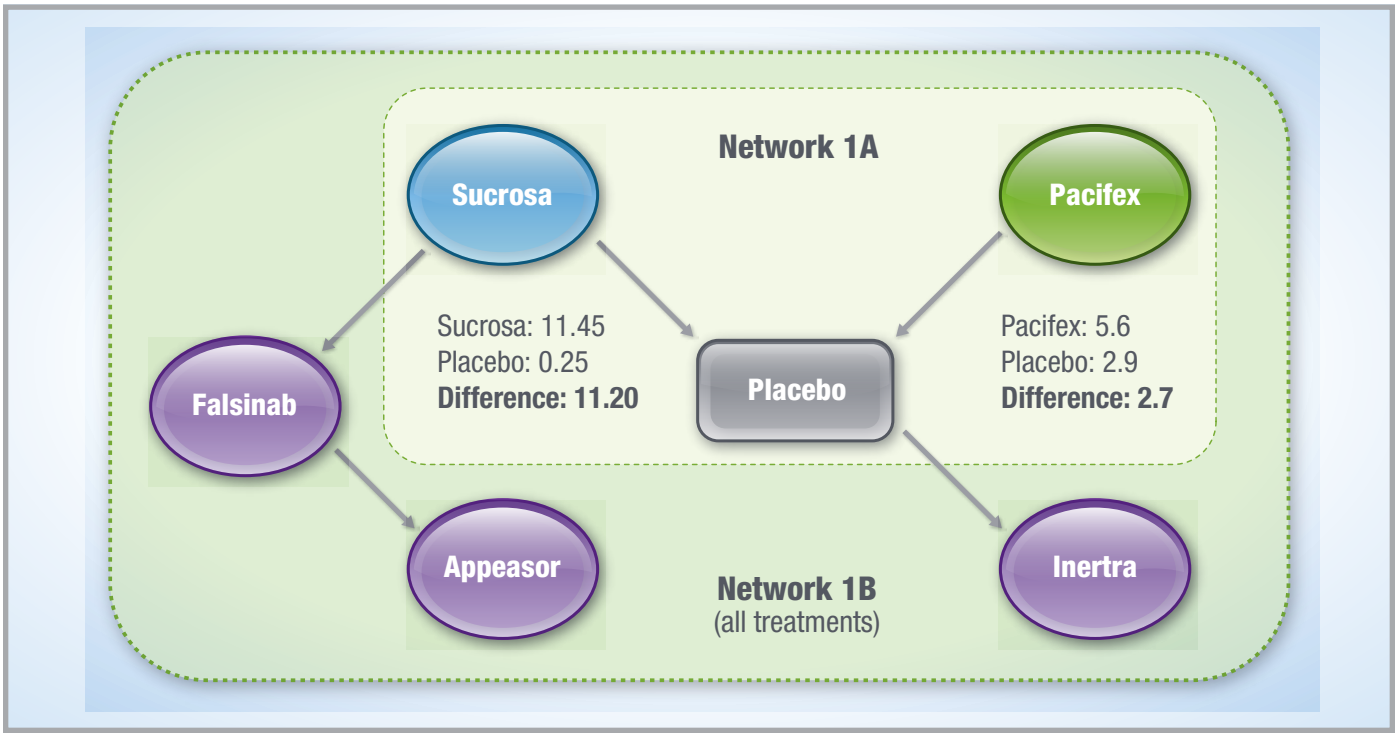


Table 2b. Estimates of Sucrosa vs. Pacifex
*(with 95% Confidence/Credible Intervals)*

| Analysis Technique | Network 1A | Network 1B (Full Network) |
|---|---|---|
| Bucher (extra decimal place in reporting) | 8.500 [4.632, 12.368] | 8.500 [4.632, 12.368] |
| Bayesian (FE) (extra decimal) | **8.502** [4.632, 12.373] | **8.499** [4.629, 12.370] |

That said, increasing the number of simulations does not make Bucher and Bayesian estimates equivalent – it simply reduces the discrepancy from trivial to even more trivial. In this context, it is worth considering what the estimates look like after we expand the treatment network to include peripheral studies (i.e., those that did not include placebo as a comparator). As shown in Table 1b, the Bucher comparison for Network 1B gives the exact same estimate as found for Network 1A, and as before, essentially the same estimate as in the Bayesian analysis.

In Table 2b, we also add an extra decimal place to the reporting; not because these numbers are meaningful (imagine, for instance, meta-analyzing blood pressure change and thinking about the third decimal place) but to re-emphasize the point that Bayesian estimates change slightly from analysis to analysis. The difference between the Bayesian results for Network 1A versus Network 1B has nothing to do with the content of the extra studies - it is simply different "Bayesian noise" at work.

The important takeaway of the Network1B results is that adding studies to the periphery does not *meaningfully* change NMA estimates. In this example, while the Falsinab vs. Sucrosa study may have information on the efficacy of Sucrosa, the study does not provide information about the *relative effect* of Sucrosa vs. Placebo, and so its addition does not change the Sucrosa vs. Pacifex estimate.

### Example 1 Takeaways

***For mean differences and hazard ratios (HRs) in simple one-study-per-comparison networks:***

- Bucher and Bayesian analyses give essentially identical results
- Bayesian results can be very slightly different depending on the number of simulations run (Bayesian results are not "exact" as they incorporate some random noise)
- Peripheral studies do not meaningfully change estimates for the treatment comparisons of primary interest

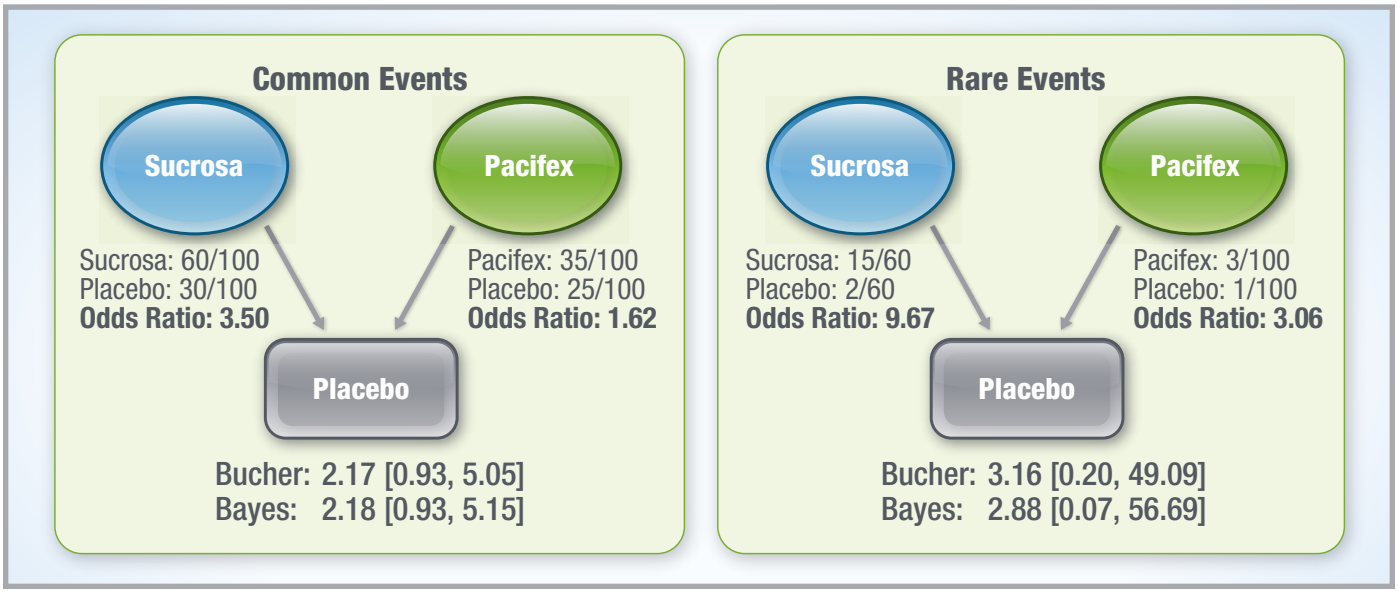
## Example 2. Discrepancies from (Slight) Modeling Differences (Odds-Ratios)

The standard Bucher and Bayesian approaches use different statistical techniques; this accounts for why they often produce similar, but not identical, results. Specifically, the Bucher method is based on a classical odds-ratio calculation, while the Bayesian approach (usually) uses arm-level data and assumes a binomial distribution to model the event rate in each arm (Figure 2 and Table 3).

Table 3. Estimates of Sucrosa vs. Pacifex
*(Odds-Ratios with 95% Confidence/Credible Intervals)*

| Analysis Technique | Common Events | Rare Events |
|---|---|---|
| Bucher | **2.17** [0.93, **5.05**] | **3.16 [0.20, 49.09]** |
| Bayesian (FE) | **2.18** [0.93, **5.15**] | **2.88 [0.07, 56.69]** |

Figure 2: Single-Study Network: Odds Ratios Examples



For common events, (i.e., where all arms have at least four events), results are only trivially different. However, in this case, the discrepancy is not primarily due to random noise and , therefore, cannot be addressed by increasing the number of simulations in the Bayesian MCMC estimation. By contrast, for rare events (roughly defined as <4 events in at least one arm), the discrepancy is often greater. This difference, however, is arguably not substantive. While a difference in odds between 3.16 and 2.88 may seem important, consider the two 95% intervals, which imply that Sucrosa may have 30 to 50 times the odds of an event compared to Falsinab, or, alternatively, perhaps only have 1/5th to 1/10th the odds. This high level of uncertainty (which would increase the more disconnected the network)[b] illustrates how indirect comparisons for rare events are *extremely* susceptible to slight differences in study methodology, event definitions, and treatment effect-modifiers (i.e., patient or study characteristics that influence treatment outcomes). The primary concern, therefore, should not be whether the Bucher or Bayesian estimates represent the "better" option but the interpretability and usefulness of the result given the wide 95% intervals.

### Example 2 Takeaways

- For some outcomes, such as odds-ratios, Bayesian and Bucher results are very similar, but not identical, due to a slight modeling difference between the two approaches.
- The differences are biggest when there are data with rare events; however, these differences pale in comparison to other issues that arise with indirect comparisons at that point.

Note that we do not need to see what would happen if we expanded the network as we did in the first example. The only change would be a miniscule difference in Bayesian results due to Bayesian noise. The peripheral studies would not affect anything else in the Sucrosa vs. Pacifex comparison.

## Example 3. Discrepancies Caused by Differences in Random-Effects Estimation

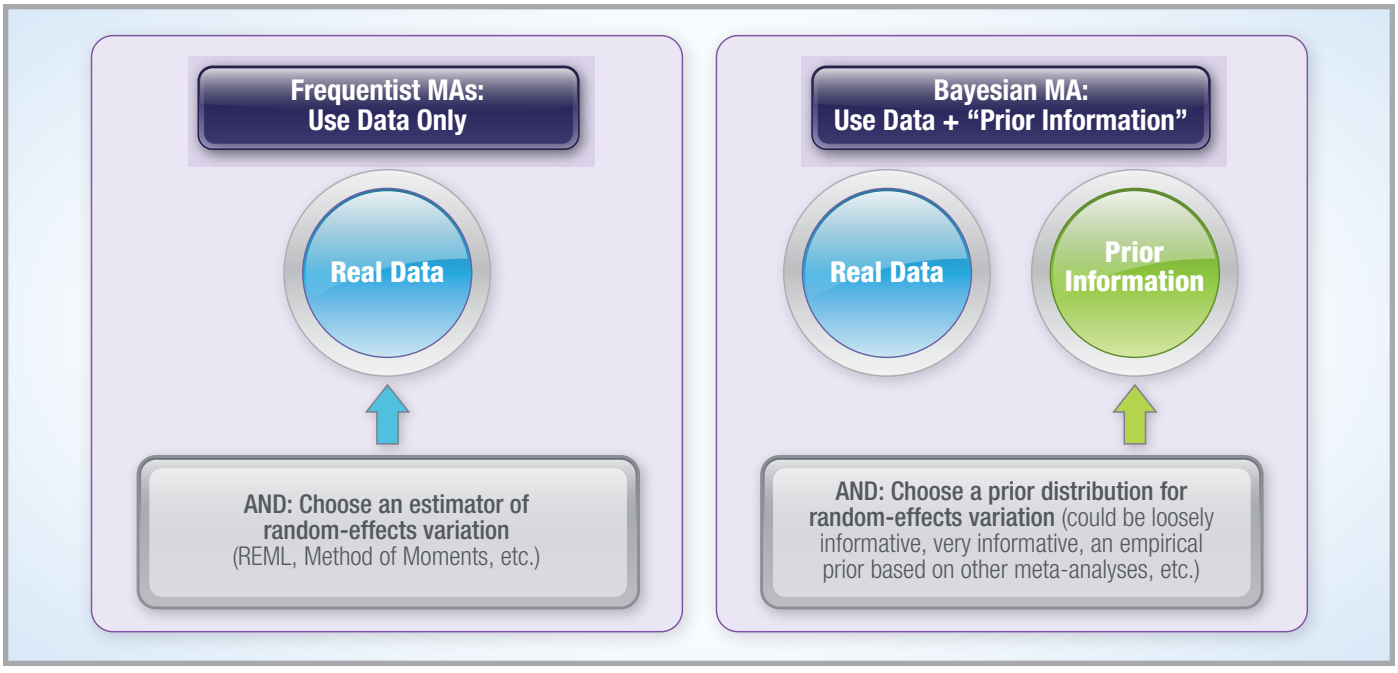
When more than one study exists for any given comparison, random-effects analyses are possible (i.e., analyses that measure and account for statistical heterogeneity – that is, variation in study effects greater than that expected from sampling error alone).

As summarized in Figure 3, Bucher random-effects analyses use classical (frequentist) random-effects meta-analyses to aggregate data for each pair-wise comparison of two or more studies, and then apply the usual Bucher calculations. If there is only one trial for a given link between treatments in the network, then the data from that study is used. Bayesian random-effects analyses start with a prior distribution for the random-effects variance and incorporate it into all estimates (Figure 4).[c]

[b] Indirect comparisons on outcomes with rare events sometimes lead to outrageously wide 95% intervals when two treatments of interest can only be compared though a long chain of studies in the evidence network. For instance, in Network 1B, an Appeasor vs. Pacifex comparison on rare events could have an upper 95% interval in the thousands.

[c] While it is not commonly done, it is possible to conduct a random-effects analysis on data from networks in which there is only a single study per comparison. In this situation, because there are no data available to help estimate a variance, the Bayesian estimate of the RE variance will be 100% dependent on whatever "prior" chosen. This might be done in situations where it is known that treatment effects tend to vary in efficacy but the dataset at hand has only one study per comparison. The only effect would be an inflation in the 95% credible intervals.

**Figure 3.** Random-Effects Bucher (Frequentist) vs. Bayesian



Before we visit our examples, however, it is important to note the difference in the "fixed vs. random-effects" choice being made for Bucher, compared to that for Bayesian.

***Bucher***
***Fixed- vs. Random-Effects Analysis***
***(as traditionally conducted)***

- Each individual meta-analysis gives its own estimate of random-effects variance, which might be zero. When it *is* zero, random-effects results are **equivalent** to fixed-effects results.
- There is no one "true" estimate of random-effects variation; different frequentist methods can give different estimates, and the "better" approach is a matter of judgment (e.g., see Veroniki 2015[1]).

***Bayesian***
***Fixed- vs. Random-Effects Analysis***
***(as traditionally conducted)***

- Random-effects results generally have (at least slightly) **wider 95% credible intervals** compared to fixed-effects results even when there is no apparent statistical heterogeneity (because we start with a prior distribution that, on average, assumes some heterogeneity).
- One global estimate of random-effects variation is used and applied to all treatment comparisons **(even single-study treatment comparisons)**
- There is no one "true" estimate of random-effects variation; different Bayesian prior distributions methods can give different estimates, and which method represents the "better" approach is a matter of judgment (see Lambert 2005,[2] Turner 2014[3])

There are three main drivers in differences between Bucher and Bayesian results.

1. The amount of heterogeneity seen in the data (e.g., none, low, moderate, high)
2. The number of studies available for each comparison (e.g., two studies available for one comparison vs. many studies available for multiple comparisons)
3. The level of variability used in the Bayesian "prior" (e.g., "zero to moderate heterogeneity" vs. "zero to high heterogeneity" vs. "zero to very high heterogeneity"). Note the last seems "safest" in that it seems to allow for the greatest range of values; however, as extremely well described by Lambert et al.,[2] such a prior also can have the effect of inflating the estimate of the random-effects (RE) variance.)

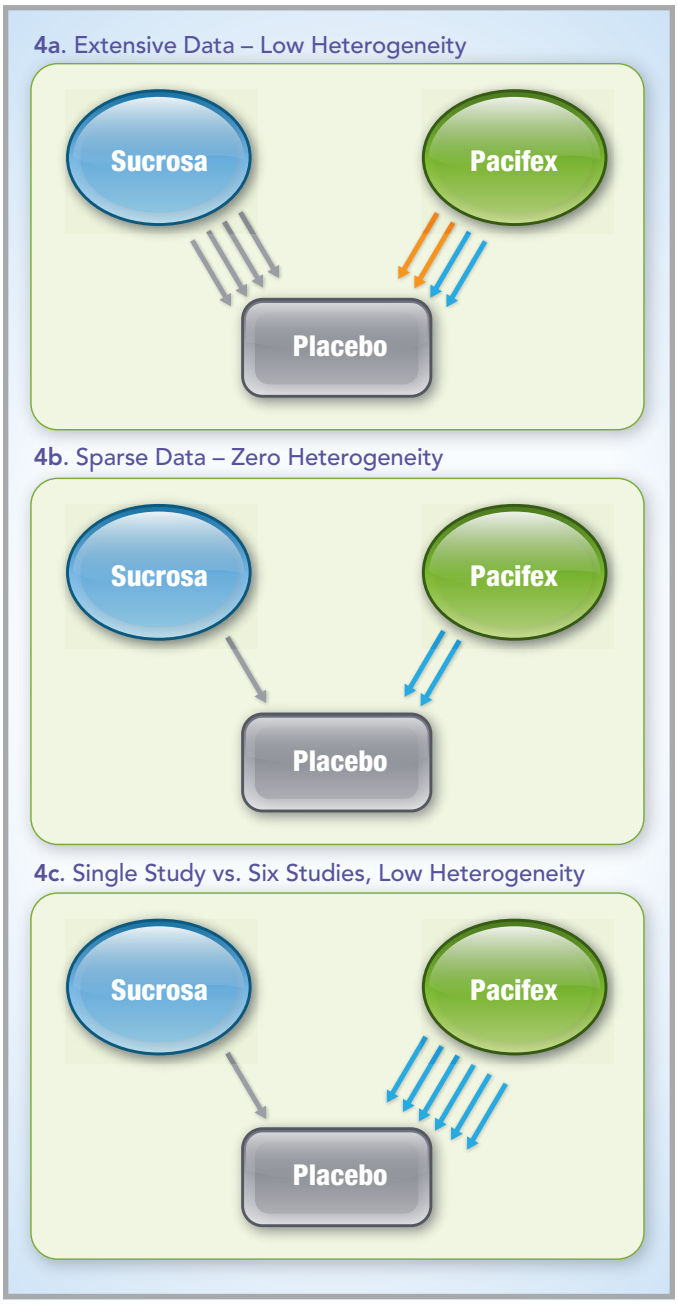
## Summary

Drivers of potential discrepancies between Bucher and Bayesian for random-effects include:

- Amount of heterogeneity
- Number of studies
- Bayesian prior used

As is obvious, adding in estimates of random-effects variation to the equation leads to a great deal of complexity in explaining potential differences between Bucher vs. Bayesian results. While we cannot cover all combinations of factors here, three exemplars will help demonstrate what sorts of differences might be expected.

Figure 4.



### Example 3a. Many Studies Per Comparison, Low Statistical Heterogeneity

Where there are many studies per comparison and low statistical heterogeneity (see Figure 4a), Bucher and Bayesian analyses result in similar estimates (see Table 3a). This is because they have similar estimates of random-effects variation. However, the 95% intervals generated by the Bayesian approach are slightly wider than those under the Bucher approach – a function of the prior used.

It is worth recalling that Bayesian results are driven by a combination of the data and the selected prior (Figure 3), so unless there is an overwhelming amount of data, the choice of prior will have some noticeable impact on the results. In this example, while our data suggest there is little heterogeneity, the Bayesian prior used here assumes (as a start) that heterogeneity is, on average, moderate or high. This prior pulls up the final estimate of random-effects variance a small amount.

*While rare, the opposite situation can also occur. If a very informative (i.e. narrow) prior distribution is used, and the variability in the observed data is *higher* than the average guess at heterogeneity represented in the prior, then the 95% interval obtained through the Bucher approach can be wider than that from a Bayesian analysis.*

### Example 3b. Two Studies for One Comparison, Zero Statistical Heterogeneity

Figure 4b presents the "worst case scenario" for discrepancies between Bucher and Bayesian results, which occurs when:

1. There is minimal information with which to estimate a random-effects (RE) variance,
2. What little information there is, suggests that there is zero RE variance, and
3. The Bayesian prior suggests that there could be a lot of RE variance.

In such a case, the Bucher approach will estimate (based on the observed data) zero RE variation, while the Bayesian approach will estimate a large amount of RE variation. This leads to much wider intervals for the Bayesian approach (see Table 3b). The Bayesian results here rely heavily on the choice of prior, since there is little observed data from which to estimate RE variation; thus, not fully trusting the observed data, the conclusion is that there most likely is a lot of RE variation.

When observing the results on the logscale, the width of the Bayesian 95% interval based on a less informative prior is almost three times the width of the Bucher interval. By comparison, when the analysis is based on a more informed prior, the interval is about one and a half times the width. The main cause of this discrepancy is as stated previously – the Bucher result is basically the fixed-effects result, while the Bayesian approach estimates substantive random-effects variance, and the width of the 95% intervals are very sensitive to the choice of prior. (The point estimates, ranging from 1.30 to 1.36, are not nearly as sensitive. The methods have different estimates of random-effects variation, and so weight the studies slightly differently, which leads to small differences in point estimates between the two analytical approaches.)

The last two examples suggest that when there are robust data and zero-to-low heterogeneity, Bayesian results tend to have *slightly* wider 95% intervals than Bucher results, and when there is sparse data, Bayesian results tend to have *much* wider 95% intervals. In cases where there is robust data and more heterogeneity, Bayesian and Bucher results are more closely aligned, as the Bayesian priors match more closely with what is seen in the data.

### Example 3c. Six Studies for One Comparison, Low Statistical Heterogeneity

Figure 4c illustrates a commonly observed network, wherein there is a well-studied (but ineffective) standard-of-care treatment (in this case, Pacifex), and a new (and believed to be more effective) treatment (Sucrosa) for which there is a single study presenting statistically significant results.

Table 3c presents the study-level data and results of an analysis that, if viewed from a manufacturer's perspective, may well prompt the following questions.

1. ***"Why isn't Sucrosa statistically significantly better than Pacifex?"*** (i.e., the HR estimated for Sucrosa vs. placebo is statistically significant, with an upper 95% confidence interval of 0.90, while the average HR for Pacifex vs. placebo is 0.93).

**Table 3a.** Estimates of Sucrosa vs. Pacifex
*(Hazard Ratios; Four Studies per Link, Low Heterogeneity)*

| Analysis Technique | Common Events |
|---|---|
| Bucher | 1.36 **[0.96, 1.93]** |
| Bayesian (wider, i.e., less informative, prior) | 1.36 **[0.86, 2.16]** |
| Bayesian (narrower, i.e., more informative, prior) | 1.36 **[0.91, 2.04]** |

**Table 3b.** Estimates of Sucrosa vs. Pacifex
*(Hazard Ratios; Two Studies for One Link, Zero Heterogeneity)*

| Analysis Technique | Common Events |
|---|---|
| Bucher | **1.30 [0.87, 1.96]** |
| Bayesian (wider, i.e., less informative, prior) | **1.36 [0.38, 5.15]** |
| Bayesian (narrower, i.e., more informative prior) | **1.34 [0.69, 2.73]** |

2. *"Why does the Bayesian estimate for our drug vs. placebo no longer look statistically significant?"* (i.e., the Bayesian 95% interval estimated for Sucrosa vs. placebo is not the same as the 95% interval reported for our trial)

The answer to the first question was touched upon previously. The results of both Bayesian and Bucher NMAs are always less precise than any individual-study result or any single meta-analysis, i.e., the 95% intervals for indirect comparisons are always wider than those for any individual direct comparison. In fact, it would not be difficult to construct a scenario in which there is a significant result vs. placebo and a non-significant result vs. Pacifex even though Pacifex performed "worse," on average, than the placebo. With indirect comparisons, it is best to focus on the size of the point estimates and the width of the 95% intervals and not on whether the intervals overlap 1.0 (or 0.0 for mean differences).

The answer to the second question has to do with the nature of Bayesian analysis. Conventionally, in this approach, one global estimate of random-effects variation is used, and applied to all comparisons in the network. In this example, the study result for Sucrosa vs. placebo gives a 95% confidence interval of the treatment effect in a specific study population, while the Bayesian analysis gives a 95% credible interval for the effect across *all* similar study populations. So, while there is no between-study heterogeneity observed for the comparison of Sucrosa vs. placebo (because there is only a single study), the (non-zero) estimate of heterogeneity for Pacifex vs. placebo is applied to the Sucrosa vs. placebo result. This leads to a wider 95% interval.

If the populations in the Sucrosa and Pacifex studies are considered clinically similar, it is realistic to believe that Pacifex vs. placebo estimate of random-effects variance is generalizable to the Sucrosa vs. placebo results. Simply put, if there is heterogeneity for the comparison of Pacifex vs. placebo, we can expect that upon further investigation of Sucrosa vs. placebo that there would be heterogeneity there as well – we just can't see it yet, as there is only the one study. This means that while the single-study result for Sucrosa vs. placebo may not overlap 1.0, the Bayesian estimate of that effect across all studies may indeed do so.

**Table 3c.** Estimates
*(Hazard-Ratios; Six Studies for One Link, Low Heterogeneity)*

| Comparison | Source/Analysis Type | Result |
|---|---|---|
| Sucrosa vs. Placebo (1 study) | Study Result | 0.60 [0.40 – 0.90] |
| | Bayesian Estimate | 0.60 [0.34 – 1.07] |
| Pacifex vs. Placebo (6 studies) | Frequentist Meta-Analysis | 0.93 [0.76 – 1.13] |
| | Bayesian Estimate | 0.93 [0.73 – 1.17] |
| **Sucrosa vs. Pacifex** | **Bucher Estimate** | 0.65 [0.41 – 1.01] |
| | **Bayesian Estimate** | 0.65 [0.34 – 1.20] |

### Example 3 Takeaways

- Bayesian and Bucher random-effects point estimates are usually very similar
- Bayesian 95% intervals are usually wider than Bucher 95% intervals
- Bayesian priors can be wide or narrow
- When these priors are averaged with the data, substantive random-effects variation may be estimated even if it is not seen (yet) in the data

## In Defense of Wider (Bayesian) Intervals

Our exploration of the source of discrepancies between the results of Bucher vs. Bayesian analyses started simply enough, with the finding that the two sets of results ranged from being "identical-within-rounding-error" to "extremely similar" to "still, pretty similar." However, once random-effects variation had to be considered, the low level of

### Bayesian Priors for Random-Effects Variation

Bayesian Estimates = Prior information + Data

- All Bayesian models start with a "first guess" for each statistical parameter. Each guess has the form of a probability distribution – the so-called *prior distribution*.
- For many parameters, data drives all estimates, and priors are truly "non-informative."
- For random-effects variation, however, the prior information chosen can have a noticeable effect. There is no such thing as a truly "non-informative" prior.
- Conventional priors for random-effects variation have a wide range (e.g., the guess is that variation is zero to "very high"), though it is increasingly common to use less vague, more informative priors (e.g., zero to "moderate").
- If the average guess at variation in Prior Information is different than what is in the Data (either higher or lower), the Bayesian Estimates will get pulled in that direction. The amount of the pull depends on how much data is available.

discrepancy between the analytical approaches held for point estimates, but not for the width of the 95% intervals. While the size of the discrepancies was heavily dependent on the number of studies available and the amount of heterogeneity in the data, another key driver for the difference was the "prior information" used in the Bayesian analyses for random-effects variance.

The argument about which of the two approaches is better varies and in some cases, is quite philosophical with regards to the applicability of estimates of variation from prior meta-analyses; the difference in interpretation of results for frequentists vs. Bayesians; the meaning of "prior knowledge"; and so on. However, from a practical standpoint, most HTA bodies see little harm in being conservative by risking an overestimation of the width of 95% intervals as opposed to risking underestimation, and they understand how poor the estimate of random-effects variation is when, for example, only two or three studies are available for a particular treatment comparison. Simply because a small number of available studies show no heterogeneity does not mean there is none, yet that simplistic implication is inherent in a Bucher ITC (indirect treatment comparison). Furthermore, it is rare for indirect comparisons to show "significant" differences (i.e., 95% intervals that do not overlap 1.0 for ratios, or 0.0 for mean differences). So generally, little is lost in basing conclusions about the treatment comparisons on the potentially more conservative 95% intervals generated by the Bayesian approach. Finally, the growing popularity of empirical prior distributions (which tend to be less conservative/more informative than the common default priors, e.g., Pullenayegum 2011,[4] Turner 2014,[3] Rhodes 2015[5]) will lead to even less of a discrepancy between Bucher and Bayesian results.

### Final Takeaways

- For single-study-per-link networks, Bucher vs. Bayesian results are near-identical
- For multiple-study-per-link networks, Bayesian results are likely more conservative (but arguably more realistic)
- There may not be much risk in 95% intervals being conservative
- Bayesian models as the base-case will offer more flexibility in general

Given that the Bayesian approach copes better with "closed-loop" evidence networks and also allows the use of meta-regression and other model additions, it is not surprising that it is the approach preferred by NICE and many other HTA bodies. But as indicated above, Bucher analyses certainly still have a place. ■


### Acknowledgments

The author would like to thank the following colleagues from the Meta Research team at Evidera for their expertise, input, and review of this article: Heather Burnett, Research Scientist; Ike Iheanacho, Research Scientist and Senior Director; Jialu Tarpey, Associate Statistician; and, Binod Neupane, Statistician.



*For more information, please contact Kyle.Fahrbach@evidera.com.*


### REFERENCES


1. Veroniki AA, Jackson D, Viechtbauer W, Bender R, Bowden J, Knapp G, Kuss O, Higgins JP, Langan D, Salanti G. Methods to Estimate the Between-Study Variance and its Uncertainty in Meta-Analysis. *Res Synth Methods*. 2016 Mar;7(1):55-79. doi: 10.1002/jrsm.1164. Epub 2015 Sep 2.
2. Lambert PC, Sutton AJ, Burton PR, Abrams KR, Jones DR. How Vague is Vague? A Simulation Study of the Impact of the Use of Vague Prior Distributions in MCMC Using WinBUGS. *Stat Med*. 2005 Aug 15;24(15):2401-28.
3. Turner RM, Jackson D, Wei Y, Thompson SG, Higgins JP. Predictive Distributions for Between-Study Heterogeneity and Simple Methods for their Application in Bayesian Meta-Analysis. *Stat Med*. 2015 Mar 15;34(6):984-98. doi: 10.1002/sim.6381. Epub 2014 Dec 5.
4. Pullenayegum EM. An Informed Reference Prior for Between-Study Heterogeneity in Meta-Analyses of Binary Outcomes. *Stat Med*. 2011 Nov 20;30(26):3082-94. doi: 10.1002/sim.4326. Epub 2011 Aug 25.
5. Rhodes KM, Turner RM, Higgins JP. Predictive Distributions were Developed for the Extent of Heterogeneity in Meta-Analyses of Continuous Outcome Data. *J Clin Epidemiol*. 2015 Jan;68(1):52-60. doi: 10.1016/j.jclinepi.2014.08.012. Epub 2014 Oct 7.