ALL-IN meta-analysis

Schure, J.A. ter

Discussion and future work

In the course of my Ph.D., I have always kept an eye on developments at Cochrane, previously known as the Cochrane Collaboration. Cochrane is an independent, international non-profit organization that is a leading authority on the methodology for systematic reviews and meta-analysis of clinical trials. It has found itself in quite some turbulence recently – e.g. dropping the "collaboration" from its branding – that has even led to insiders asking: "Has Cochrane lost its way?" (Newman, 2019). Part of this criticism came from members that felt that Cochrane should not only be the authority on systematic reviews but should also lead the way in improving the primary evidence: improve how and when clinical trials are performed.

While new procedures were implemented at Cochrane on living systematic reviews and network meta-analysis, not much changed in the basic statistical recommendations. In this discussion, I would like to reflect on how ALL-IN meta-analysis relates to standards at Cochrane in updating meta-analyses and judging redundant trials in cumulative meta-analysis. The concluding section discusses future work.

Updating meta-analyses

In 2018, a scientific expert panel was asked: "Should Cochrane apply error-adjustment methods when conducting repeated meta-analyses?" Its answer was "no", so Cochrane meta-analysts could simply continue their practice of recalculating p-values and confidence intervals each time a review was updated. Obviously, the expert panel knew that this practice increases type-I errors. For meta-analyses specifically, the false-positive risk of updating meta-analyses is estimated to range between 10% and 30% (Borm and Donders, 2009; Imberger et al., 2016); a lot more than the 5% that meta-analysis usually sets out for. In my view, the recommendation to stick to basic statistical methods had a lot to do with practical limitations of the available methodology at the time. These limitations do not apply for or can be mitigated by ALL-IN meta-analysis.

A year earlier Simmonds et al. (2017) had provided a review of all possible methods available for sequential meta-analysis that was part of the expert panel's deliberations. Two practical arguments dominate this review and the discussion of the expert panel (Cochrane Scientific Committee et al., 2018). The first is the lack of control over the primary studies. The second is the need to model heterogeneous results.

The meta-analysis has no control over primary studies

Most approaches discussed by Simmonds et al. (2017) take methodology from sequential clinical trials – either group sequential methods or its generalization in α-spending (Lan and DeMets, 1983)– and apply these to meta-analysis. What is lacking in meta-analysis (but exists in clinical trials) is control, and that lack of control is the main challenge of applying these methods outside a single clinical trial. Statistical properties for error control rely on stopping rules, and only if those can be enforced does the methodology guarantee type-I error control. Both group sequential methods and α-spending set a maximum sample size or a maximum number of looks and a spending strategy for α that originated with either Pocock (1977) or O'Brien and Fleming (1979). This defines the threshold for the Z-statistic and guarantees the 5% α type-I error if the data collection stops either when the threshold is crossed or we arrive at the maximum sample size or number of looks. The unfortunate consequence is that, strictly speaking, the results become uninterpretable when these thresholds and sample size cannot be enforced and might be violated.

Moreover, if accumulation bias processes are at play, the disagreement between the stopping threshold and actual stopping might be more pronounced. A first meta-analysis, e.g. on two studies, depends on the results so far if there is a chance that the first single study result would have halted any further studies. Without any meta-analysis result available, clinical trials might already know of each other's results and apply implicit stopping rules or accumulation processes. This invalidates the meta-analysis stopping rule as soon as clinical trialists have an evidence-based reason to carry out their trial or not – as they should have.

This is what sets ALL-IN meta-analysis apart: Ville's inequality is stopping-rule-free[2]. No process that decides to stop the meta-analysis, decide its *timing*, and no process that decides the accumulation of the underlying studies can invalidate its results.

Modelling heterogeneous results

The methods in the Simmonds et al. (2017) review try to capture random-effects meta-analysis by including a measure of between-trial variability. This heterogeneity parameter is difficult to estimate over time. Including a study in the meta-analysis that is very different from the earlier ones can increase the between-trial variance estimate and as such decrease the effective sample so far. This leads to strange behavior and the observation by Kulinskaya and Wood (2014) that sequential meta-analysis can be better off when many small trials are included than if a few very large trials are. This disagrees with the general notion of quality in clinical trial research that prefers large over small trials.

Any random-effects methodology for ALL-IN meta-analysis will have to deal with the same issues. My recommendation, for now, is therefore to use close collaboration as a tool to decrease heterogeneity (Section 1.3). This is in agreement with the recommendation from

[2] This property is shared by one other method in the Simmonds et al. (2017) review: a proposal to use the law of the iterated logarithm. There are close connections between this approach and ALL-IN meta-analysis in the work of (Robbins, 1970). The specific proposal discussed for meta-analysis has the disadvantage of requiring some constants to be set that are not very intuitive.

Tierney et al. (2021) to align the objective and eligibility criteria. This requires more of a cultural change than a statistical one, however.

It is comforting that also Richard Peto believed that systematic reviews should not be bothered too much by heterogeneity. According to Senn (2000) he even straight-out opposed random-effects modeling. His own words are: “In performing overviews, we are not trying to provide exact quantitative estimates of percentage risk reductions in some precisely defined population of patients. We are simply trying to determine whether or not some type of treatment tested in a wide range of trials produces any effect on mortality” (Peto, 1987).

Representativeness is part of a recurring discussion in the clinical trial methodology literature. Many statisticians and methodologists oppose the view that trial estimates are representations of some population effect. The most colorful viewpoint that I found is by Rothman et al. (2013), which mocks calls for more representativeness in trials as “exacted along with motherhood apple pie and statistical significance”. They agree with Peto that it is not that important. The main aim of clinical trials is to construct general statements – controlling confounding variables and understanding causal mechanisms – instead of estimating a population effect. “It is not representativeness of the study subject that enhances the generalization, it is knowledge of specific conditions and an understanding of mechanisms that makes for a proper generalization.” (Rothman et al., 2013) If we believe that the trials we include study a causal mechanism well, then their fixed-effects meta-analysis estimate can be used to evaluate the uncertainty and update our current evidence-base.

Redundant trials in cumulative meta-analysis

The term *cumulative meta-analysis* refers to applying meta-analysis to a growing series of studies, usually by using no other methods than any conventional meta-analysis would. Baum et al. (1981) seem to be the first to do this, but the term is introduced by Lau et al. (1992), describing its rationale as follows: “Performing a new meta-analysis whenever the results of a new trial of a particular therapy are published permits the study of trends in efficacy and makes it possible to determine when a new treatment appears to be significantly effective or deleterious.”

Many of such cumulative meta-analyses are performed retrospectively, to judge in which year trial data could have reached a conclusion and no further trials should have been performed. The approach was also immediately criticized, however, for applying single sample-size confidence intervals, uncorrected for multiple looks, to repeatedly test the same null hypothesis (Lau et al., 1995). There is an increasing interest in studying the “redundancy” of trials in such ways, as the Evidence-Based Research Network presented at their second conference (Evbres, 2021). They found that 31 studies performed some sort of cumulative meta-analysis between 1981 and 2021. These do not agree on how to judge redundancy, however. While most of these cumulative meta-analyses used a statistical threshold in their sample to decide when the sufficient trials ended and the redundant trials began, they managed to use 10 different ones!

ALL-IN meta-analysis can be used to judge new trials as “redundant” in two ways: for efficacy and for futility. For efficacy, the threshold $1/\alpha$ can be used that relies on both a pre-set level of α and a pre-set effect size of minimal interest. If those are available, there is only one threshold for the Z-score that can be used to decide whether further trials are redundant. This means that one can decide redundancy based on demonstrated efficacy but not based on futility. To deal with futility as well, confidence sequences can be used that are anytime-valid (Section 1.1.5, Section 2.4.2). This approach distinguishes ALL-IN meta-analysis from the Wald sequential probability ratio test (SPRT) that has a lower threshold for futility that needs to be enforced to guarantee the properties of the upper threshold for efficacy. If the sequence of confidence intervals is closing in on very small effects, and the interval contains only parameter values that are smaller (closer to the null) than the effect of minimal interest, the line of research can still be considered futile. This is an intuitive notion of futility and a straightforward decision if a pre-set effect of minimal interest is available. In that case, the meta-analysis can advise against more (redundant) trials.

The meta-analysis has little control over what happens next, however. There is always the possibility that somewhere around the world a new trial is started. Even after a boundary is reached for efficacy or futility, we want the meta-analysis to give the most complete synthesis of the evidence base and include the new trials. For ALL-IN meta-analysis this is no problem since the e-value and confidence intervals can still be updated. Fortunately, in the fixed-effects meta-analysis presented in this dissertation, the uncertainty can never increase. The intervals can only shrink (for a running intersection confidence sequence) and the e-value can never undo a rejection of the null hypothesis (once the threshold is reached the decision to reject has type-I error control). So even if the meta-analysis is concluded and any new trials considered redundant, there is the possibility to extend the meta-analysis and give a complete evaluation of the evidence. This evaluation can supplement, but not undo, an earlier decision for redundancy that has error control for rejecting a null hypothesis (for efficacy) or rejecting an effect of minimal interest (for futility).

Future work

Meta-analysis beyond summary statistics

ALL-IN meta-analysis is ready to be applied to summary statistics if they construct valid Z-statistics. As we write in Chapter 6, however, I agree with Lawrence et al. (2021) that meta-analysis on the raw trial data – so-called IPD-meta-analysis, for Individual Patient Data – would serve science much better. Statistical methods for IPD-ALL-IN meta-analysis are partly available and partly still under development. Turner et al. (2021) introduces e-values and confidence sequences for 2x2-tables, that can be easily generalized to an anytime-valid version of the Cochran-Mantel-Haenszel test in meta-analysis. Also for time-to-event data, we are developing confidence sequences for the hazard ratio that improve on the Peto estimator if IPD-meta-analysis is possible. In general, methods for regression, like linear regression and the Cox model, are a major goal for future work. Another very interesting direction of future research is to combine ALL-IN meta-analysis with network meta-analysis, where there is also an interest in correct inference after updating the meta-analysis (Simmonds et al., 2017).

Error control for the pseudo-Bayes posterior odds

The notion of pseudo-Bayes posterior odds in Chapter 5 and its appendices needs further development. It might not be so easy to combine the notion of *safety* (Grünwald et al. (2019), Theorem 2.0.2: for all $P \in H_0$ $\mathbf{E}_P(\mathrm{BF}^{\mathrm{ps}}) \leq 1$) with error control for the pseudo-Bayes posterior odds. We expect that accumulation processes or stopping rules exist for which the latter does not hold.

We would like to connect this research to other work on the usefulness of Bayes factor calibration (De Heide and Grünwald, 2021) and the difference with Bayesian paradoxes that are more like publication bias than accumulation bias (Dawid, 1994; Senn, 2008).

Data sharing and rank tests

Chapter 6 raises questions about the necessity of data transfer agreements in a live meta-analysis like ALL-IN-META-BCG-CORONA. I plan to write a paper with a lawyer as my co-author that answers these questions to guide future live meta-analyses. The privacy sensitivity of this particular meta-analysis lies in the dates at which participants enter the study (are randomized to either placebo or vaccine) and the dates at which they experience Covid-19 infections and/or are hospitalized with Covid-19. These calendar dates are important because we analyzed this particular meta-analysis on a calendar time scale.

Left-truncation and staggered entry If participants do not all enter the study at once, one of two things happens that I – following the literature – will call ‘left-truncation’ and ‘staggered entry’. Whether our analysis has to deal with either of the two depends on the chosen time scale most relevant to the occurrence of the events[3].

On the one hand, time-to-event can be calendar time, e.g. time to an infection that occurs in (epidemic) waves. All participants in a risk set share a hazard if they are in follow-up and event-free on the same calendar date, such that late entry occurs as left-truncated event times. Left-truncation means that participants only enter the risk set once they enter the study but have already ‘survived’ some calendar time that might have observed an event for other participants. Nevertheless, they should not be part of the risk set to evaluate events that happened before they entered, since we know that an event before entry is impossible, e.g. because being alive or more general event-free is an inclusion criterion for study enrollment.

On the other hand, time-to-event can be participant time, specific to each participant, e.g. time since surgery. All participants in a risk set of an event share a hazard if they are in follow-up and event-free for the same time since their own specific date of enrollment/randomization/intervention, such that late entry occurs as ‘staggered entry’. Staggered

[3] This explanation (this exact wording) also appears in two tutorials I wrote on left-truncation and staggered entry that are available on our SafeStats and All-IN meta-analysis project page (Ter Schure et al., 2020a).

entry means that participants that enter late could still enter the risk set of events that happened earlier, for events of participants that had the same participant time since their own date of intervention, as the late entered participant experienced since its date of intervention.

Hence in a 'left-truncation' analysis, participants that enter late can only enter the risk set of events that happen after (in calendar time) they enter the study, while in 'staggered entry' analysis, participants that enter late can enter the risk set of events that already happened. Left-truncation is no problem for our safe logrank test. Staggered entry, on the other hand, breaks the independent increments property that we need for the underlying martingales – those that drive our anytime-valid analysis.

In Chapter 2 we do not recommend using our safe logrank test under staggered entry. Other sequential logrank tests, however, might suffer from the lack of an appropriate martingale just as much. For the logrank statistic the literature shows that asymptotic results are hopeful (Sellke and Siegmund, 1983), as long as certain scenarios are excluded (Slud, 1984). I wonder how valid these results remain for small studies (e.g. surgery trials) with severe staggered entry.

Rank tests In studying the staggered entry problem, my colleague Muriel F. Pérez-Ortiz thought of an exact rank test that does construct a martingale under staggered entry. Without staggered entry, it is very similar to the logrank test, but with staggered entry, it is quite different. In future research we hope to investigate how powerful this test is, and if it is not, whether there are scenarios with severe staggered entry that would make the use of this test appropriate.

I can already think of one such scenario: live meta-analysis with easy data sharing. A pure rank test means that trials only have to share rank data, which is minimal in terms of privacy risk. Live analysis of ranks means that they share the ranks by calendar date. At each calendar date with an event, we need to know the group in which it occurs – treatment or placebo – and where that event ranks in time-since-randomization in comparison to the earlier events. We do not have to know what the event time was, or what the calendar date of randomization was for the participant that experienced the event. So the meta-analysis statistician cannot recognize any participants based on their date of entering the trial or how long it has been since their randomization. If you recognize a participant by the date of their event alone (the date of their rank), you probably also already knew that the person was in the trial.

Thresholds

If e-value research aims to serve Evidence-Based Research it is very interesting to look into the various thresholds already used to decide on redundancy in cumulative meta-analysis. Of course, many of them will not be statistically valid, but they might give more insight into what users of statistics expect from their methods and help us improve our communication of what e-values can do. Maybe some will only consider efficacy, while others also consider futility. Maybe some of them are inspired by Bayesian reasoning (Chapter 5), while others are more frequentist (Chapter 4). What matters the most to those that worry about clinical trial priority setting (Chalmers et al., 2014) might help set the priorities for the statisticians working on anytime-valid inference.

Statistical communication

This dissertation started with p-values and I would like to go full circle and conclude it with p-values as well. One major inspiration for my work on e-values is that p-values are so often misunderstood (Gigerenzer, 2018; McShane and Gal, 2017). I have good hopes that we can improve on that if we teach statistics with more reference to gambling. Personally, I find the scale of betting scores much more intuitive than that of p-values; yet I have no empirical evidence that statistical beginners would think so as well. In Ter Schure (2021c) I propose to design an experiment to test this hypothesis; at least find some evidence against the idea that both p-values and betting are both simply *too difficult*. I still want to do that and – with the help of Daniel Lakens – have good hopes that we can start with a pilot experiment. His open Coursera courses already show that many statistical beginners do want to understand what is going on with statistical testing.

Without having played a real poker game or entered a casino, I feel that the mathematics of strategic gambling is exciting. I hope that ALL-IN meta-analysis can encourage that excitement in others, increase enthusiasm for statistics, and help meta-analysts recognize the crucial role they play in strategic science. “Standing on the shoulders of giants.”

Judith ter Schure
Utrecht, November 2nd, 2021