ALL-IN meta-analysis

Schure, J.A. ter

6 | Data sharing in a live meta-analysis

The scientific response to the Covid-19 pandemic was far from perfect. Conflicting results on hydroxychloroquine made officials in the U.S. first recommend the anti-malaria drug and then warn against it. Similarly, systematic reviews on ivermectin could not draw robust conclusions when they first included results in the meta-analysis and had to exclude them later after their retraction from a preprint server. How different was the response of the research community studying the Bacillus Calmette-Guérin (BCG) vaccine! No BCG researcher went on television to state that they single-handedly proved that the BCG vaccine – originally developed to protect against tuberculosis – makes us invincible. On the contrary, the BCG community worked closely together and remained cautious until this day. The results will be published later this year, so here we want to simply chronicle how it all started and what we learned along the way.

Early 2020, BCG researchers from the university medical centers of Utrecht and Nijmegen were among the first to announce their clinical trial (newspaper Trouw, Van der Wier, 2020, March 18). Not only were they early, but also generous in sharing their protocol when other researchers around the world started similar trials. Already from the beginning these trials had much in common and great potential to be analyzed together. The chaos surrounding hydroxychloroquine shows how important coordination can be. The story of ivermectin illustrates the risks of a meta-analysis that waits for summary estimates to appear in (preprint) publications.

Even if trials are performed perfectly, however, unreliable results can arise due to multiple testing when many trials address the same question simultaneously. Fortunately, the BCG researchers were warned of this risk by their trial statistician dr. Henri van Werkhoven. A consequence of this risk is that the first trial to find an effect could be an outlier, but still be published quickly and threaten the continuation of the other trials. In the urgency of a pandemic, a disagreeing meta-analysis might come too late to start the trials up again.

ALL-IN meta-analysis

When we offered the Dutch BCG researchers a solution to this problem they had already contacted many of the trials around the world. For statistical validity, the BCG trials needed coordination and needed to be analyzed together. For efficiency's sake, we should start the statistical analyses as soon as possible. Our plans got a name: ALL-IN meta-analysis, for Anytime Live and Leading Interim meta-analysis. We provided the statistical methodology to analyze all these BCG trials together continuously, while they were still ongoing.

The BCG researchers courageously embraced our novel methods and ALL-IN-META-BCG-CORONA was born (Van Werkhoven et al., 2021). It became a collaboration by two groups of clinical studies, of 7 and 4 each, that decided on trial selection together, shared their data at interim stages, and monitored the results live in a dashboard. We will focus here on the 7 trials that studied healthcare workers (the others study the elderly) and on the outcome measure of Covid-19 infections (the other being severe Covid-19 infections requiring hospitalization).

The main goal was to find out whether an immune response to BCG provides indirect protection against Covid-19. If a beneficial effect were to be confirmed quickly, this could save many lives since BCG is widely available around the world, which was not the case for any other treatment or Covid-19 specific vaccine at the time. On the other hand, if futility or harm could be confirmed, studies could be stopped early and resources saved and put to better use elsewhere in the scientific response to the pandemic.

Lessons learned

Working in a large-scale multi-center collaboration across the globe comes with many lessons. First, we learned the intricacies of time in sequential time-to-event analysis in our development of the safe logrank test and anytime-valid confidence sequences for the hazard ratio (Chapter 2). Second, we realized the need for a software package safestats (Turner et al., 2022) with transparent tutorials and a webinar (Ter Schure et al., 2020a). Third, we were confronted with meta-analysis issues that arise from a bottom-up collaboration, like heterogeneity in trials, (dis-)agreement on decision rules, and interpretation of results. The experience does make us hopeful about the benefits of the approach in general, in terms of efficiency, collaboration, and communication (Chapter 1). Finally, we learned about data sharing, which is what we would like to discuss here. We came up with solutions to the issues at hand, but still have questions to discuss that remain.

Crucial is that the statistical test and confidence intervals that we developed and applied are valid at any time. Figure 6.1 shows the dashboard that we used to communicate interim results to the participating trials. (The dashboard is in a demo mode and based on synthetic data: "fake" values for each trial based on public trial characteristics.)

We kept track of an e-value (Grünwald et al., 2019), a measure of evidence and test

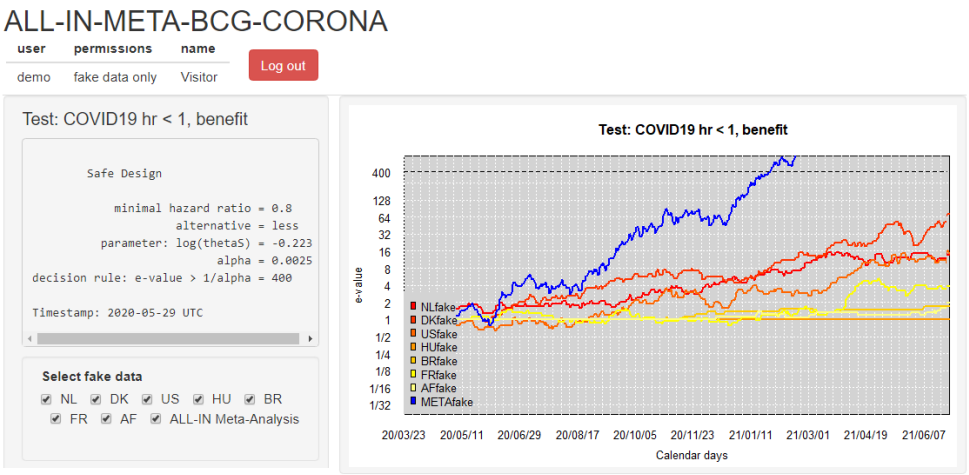

Figure 6.1. *Dashboard used to communicate interim results in ALL-IN-META-BCG-CORONA to all data uploaders with a login. The involved trials were performed in the Netherlands (NL), Denmark (DK), the United States (US), Hungary (HU), Brazil (BR), France (FR), and Guinea-Bissau/Mozambique (AF). The dashboard is in demo mode with "fake" values. Note that the y-axis is on the log scale.*

statistic that we compared to the threshold $1/\alpha = 400$[1]. Whenever the cumulative meta-analysis e-value would cross this threshold, we could declare statistical significance – you can think of an e-value as $1/p$-value, with the crucial difference that it keeps its validity for testing irrespective of when we stop the data collection. This procedure guarantees type-I error control at level α=0.0025 regardless of the sampling plan, the number of analyses, or their timing. Apart from hypothesis testing, we also kept track of confidence intervals that were anytime-valid. In Figure 6.3 we show examples of these.

Practical hurdles arise when data are shared. This is true in general, but in our ambition to do a live analysis this was even more pronounced. We wished to retrospectively process each newly updated trial data set to show how the evidence since the last upload had changed. Not only to find a conclusion of benefit as early as possible, but we also wanted to show whether the evidence was moving in the right direction and make it possible to prepare for future conclusions. By showing an e-value for each calendar day, our dashboard allowed users to spot trends in the evidence very easily.

[1]This level of α agrees with the FDA's two trial rule (two trials at level $\alpha = 0.05$ give a total level of $0.05^2 = 0.0025$), but was argued by attributing 10% of $\alpha = 0.05$ to the outcome measure of Covid-19 infections and 90% to a co-primary outcome measure of severe Covid-19 infections requiring hospitalization (with very few occurrences in a population of working-age healthcare workers).

6.1 Sharing live results while keeping researchers blinded

In clinical trials like the BCG trials, most involved researchers and doctors are blinded to the allocation of treatment and placebo, as well as to any results. After all, if you know that the results point towards effective treatment, this might also indicate whether a participant that is improving has received treatment or placebo. For our analysis, however, we needed at least one person to handle a trial's data fully unblinded. This turned out to be no problem since most trials had a trial statistician that would also perform interim analyses and/or provide interim data to a data safety and monitoring board. We asked for this person to be the data uploader for the meta-analysis.

These data-uploaders were the first the get access to the dashboard, each with personal login details. The reason is that the dashboard could reveal a participant's allocation to those that still needed to remain blinded. The dashboard of Figure 6.1 shows that the line goes up if an event occurs in the control group (evidence against the null brings us closer to the threshold at 400 for benefit) and that the line goes down if an event occurs in the BCG group. This level of detail in the dashboard reveals both the time and place of occurrences of Covid-19 by the calendar date and trial. If you observe that a sequence of e-values goes up at a certain calendar date, and you know the person that tested positive for Covid-19 that day, you can deduce with certainty that the person was randomized to placebo (and similarly to BCG if the e-values go down).

In the early stages of the meta-analysis, these logins only gave permission to view the overall meta-analysis e-values and the data uploader's own trial contribution. This mechanism made sure that no one could access the dashboard that needed to stay blinded to interim results and that the data uploaders could not access privacy-sensitive data from other trials. Observed Covid-19 infections from other trials were bundled together in the meta-analysis e-values such that the location of those events could not be derived from the dashboard.

Remaining questions

- If we would group more than one observation of Covid-19 by calculating an e-value by week or month, instead of by day, would that make it sufficiently hard to deduce randomization from observed events? Would that be enough to allow data-uploaders (not blinded to their own trial results) to inspect other trials' e-values? Or even to allow all participating researchers (blinded to their own trial results) to inspect all trial results except their own?

6.2 A central analysis

In collecting the data from the trials we had two options for analysis by calendar date (as in the dashboard in Figure 6.1). Either do a meta-analysis based on summary statistics (so-called two-stage meta-analysis) or do a meta-analysis on the raw data (so-called IPD meta-analysis, for Individual Patient Data). While this decision is a familiar one in meta-analysis, for the first option, we had to ask the data-uploaders something completely

unfamiliar. We did not only want them to share new summary statistics at each data upload but to share a sequence of summary statistics by calendar date each time they uploaded new data. On the other hand, for the IPD-analysis of the second option, there was nothing special and we needed all trials to simply upload their data so far to an upload-only folder that only we could access. We chose that second option.

Figure 6.2 shows the type of data we requested for our BCG analysis. The analysis was stratified by hospital, so for each healthcare worker randomized in the trial, we had to receive information about their location. We allowed the trials to label the hospitals ('A', 'B') without actually naming them since by knowing the hospitals, the data would become more privacy-sensitive. If you know the hospital and the calendar date that someone entered the study (dateRand), you could recognize that person in the raw data and identify whether the person had Covid-19. The approach did not mitigate this risk entirely, though, since some trials were performed in a single hospital so their participants could still be recognized in this way.

intervention	dateRand	hospital	COV19	dateCOV19	COV19hosp	dateCOV19hosp	dateLastFup
control	2020-05-07	A	yes	2020-05-11	yes	2020-05-15	2020-06-23
control	2020-05-04	B	yes	2020-05-08	yes	2020-05-12	2020-06-23
BCG	2020-05-08	A	yes	2020-05-21	yes	2020-06-01	2020-06-23
control	2020-05-07	B	yes	2020-05-25	no	NA	2020-06-23
BCG	2020-05-05	A	yes	2020-05-24	no	NA	2020-06-23
BCG	2020-05-10	B	yes	2020-06-03	no	NA	2020-06-23
control	2020-05-14	A	yes	2020-06-23	no	NA	2020-06-23
control	2020-05-10	B	no	NA	no	NA	2020-06-23
BCG	2020-05-08	A	no	NA	no	NA	2020-06-23
BCG	2020-05-04	B	no	NA	no	NA	2020-06-23

Figure 6.2. *Example (fake) data set from the working instructions to data-uploaders (Ter Schure et al., 2020a).*

Remaining questions

- Would it even be possible to collect summary statistics by calendar date from trials? Maybe if we would write an R script that each data-uploader could run locally? Or would too many problems arise for the time we had to overcome them and would this not be any faster than sharing the raw data?
- Would it even be possible to ask all involved trials to work in R? By allowing the use of other software packages, we risk that trials share incorrect summary statistics. What we asked for was not trivial, since we analyze the data as left-truncated calendar time. Even in R, no standard software outputs the right logrank statistic and we had to write our own.

6.3 Data transfer agreements

To share clinical trial data such as in Figure 6.2, the usual approach involves agreeing on a Data Transfer Agreement (DTA) and signing it. These agreements protect the privacy of the participants in the trial but also cause an enormous delay. For some of the trials in our meta-analysis, it took months for the lawyers on both ends to agree on the terms in the DTA. Interestingly, halfway through this process, a lawyer commented that we might not even have needed DTAs for data of the structure described in Figure 6.2.

Remaining questions

- Were these DTAs really necessary given the limited amount of data we asked for (Figure 6.2)?
- How does our requested data compare to full Kaplan-Meier plots or Epi curves, which are routinely included in medical publications?
- How do we convince trials in the future to share limited raw data without DTAs?

6.4 Estimation

So far, we have focused our discussion on testing the null hypothesis. This was the main aim in our ALL-IN meta-analysis: rejecting the null hypothesis in favor of an alternative hypothesis of minimal clinical relevance set at a hazard ratio of 0.8 (see Safe design in Figure 6.1). Such a rejection could lead to the conclusion of the meta-analysis and advice to stop the individual trials. A second aim is of course estimation. Here, two more disadvantages arise for meta-analysis on summary statistics that we, fortunately, did not encounter since we analyzed the raw data. First, a meta-analysis of time-to-event summary statistics is biased. This is a technical point that we will not discuss in detail here. For practical purposes, it is common to settle for biased estimates (Simmonds et al., 2011). For our purposes, however, there is a second disadvantage of summary statistics. If we cannot collect the summary statistics for each calendar day, they produce wider confidence intervals.

Remember that a (1-α)-confidence interval is a collection of parameter values that, if taken as the null hypothesis, each cannot be rejected at level α. This means that if we have a sequence of such confidence intervals, each of which is valid at any time, we can take its running intersection. Once a value for the parameter is rejected by an anytime-valid test, it never has to be re-included in the interval. A running intersection often achieves a narrower interval, as is shown in Figure 6.3. Making full use of this running intersection is only possible if we have can calculate the interval at each calendar date, and not if we only have limited summary statistics.

6.5 Conclusion

In summary, we believe that it was wise to not only collect summary statistics. Summary statistics are known to be much more prone to mistakes and manipulation than IPD meta-analysis (Lawrence et al., 2021). Collecting the raw data indeed allowed us to turn data

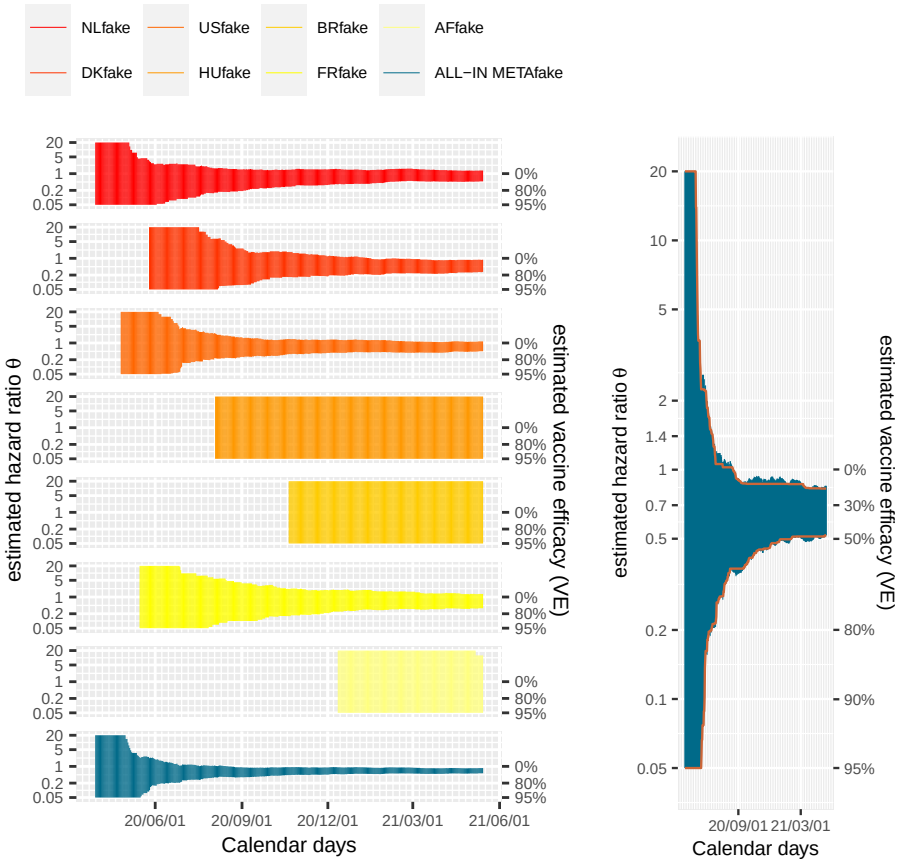

Figure 6.3. Anytime-valid confidence sequences corresponding to the fake data in Figure 6.1 with the running intersection for the meta-analysis sequence. Data generated according to the meta-analysis design with effect just slightly larger (hazard ratio = 0.7) than that of minimal interest (hazard ratio = 0.8).

cleaning into a collective effort between the data uploader and the meta-trial statistician and to spot the inadvertent mistakes. We could also confirm a suspicion of insufficient randomization in one trial which led to its exclusion due to increased risk of bias.

The main question is still whether we could have done the same approach without the delay of data transfer agreements. Crucial here is maybe how different this approach was to usual research. The involved trials put the research line before their own publication. If this becomes more commonplace, we might also view the need for DTAs very differently. Or the opposite is true and DTAs can serve a role in this transition from a science of individual interests to a science of live collaboration.

ALL-IN meta-analysis asks for a culture change that leaves behind the uneasiness of sharing your 'gold' before your own publication. ALL-IN-META-BCG-CORONA shows that this is possible in a pandemic and, hopefully, also is outside a pandemic. No one tried to make the headlines with their own study. The involved trials put collaboration before their own interests. These are the attitudes we need to increase value and reduce research waste.