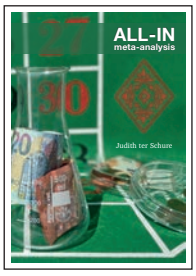
In de verdediging

| In defence



ALL-IN meta-analysis
Judith ter Schure

In April 2022, Judith ter Schure from CWI successfully defended her PhD thesis at Leiden University with the title *ALL-IN meta-analysis*. Judith carried out her research under the supervision of prof. dr. Peter D. Grünwald (UL and CWI) and dr. Daniël Lakens (TU/e).

Accumulation of knowledge

Learning in science can be considered a cumulative process. Over time, researchers perform studies, which ideally leads to the accumulation of knowledge. In clinical research, for example, often multiple studies are performed over time to measure the impact of some medicine. Promising trials may motivate researchers to do more research, and at some time a conclusion is drawn after which no new trials are deemed necessary. It is thus important that existing studies actively steer the decisions on new research. To support the decision-making process, researchers perform systematic reviews that construct a complete collection of the results of all publications that try to answer a similar question. These publications are evaluated based on quality, such that the overview gives a good impression of what is known so far. A meta-analysis adds to that with a statistical summary of the results, usually relying on hypothesis testing, confidence intervals, or p-values. Hence an important question arises: how do you statistically combine results from studies that accumulate over time? This question is not so simple to answer.

Conventional statistical methods that rely on p-values perform very poorly in meta-analyses where *time* is involved. Judith developed ALL-IN meta-analysis, a way to synthesize in one statistical analysis results obtained in multiple studies, while the collection of studies is still growing. ALL-IN meta-analysis stands for Anytime, Live, and Leading INterim meta-analysis. It provides the statistical methodology for a meta-analysis that can be updated at *any time* — reanalyzing after each new observation while retaining type-I error guarantees (probability of a false positive), is *live* — no need to prespecify the exact size of the collection of studies or the timing of the analysis, and can be *leading* — in the decisions on whether individual studies should be initiated, stopped or expanded. Let's see how this works.

p- and e-values

In statistical testing, the p-value is a notion of surprise; it tells us something about how unexpected the observed results would be assuming the null hypothesis is true. A very small p-value occurs if it would be very unlikely to observe such an extreme (or more extreme) observation by mere chance. The null hypothesis is rejected when p is less than a significance value α, conventionally chosen equal to 0.05.

Pas gepromoveerden brengen hun werk onder de aandacht. Heeft u tips voor deze rubriek of bent u zelf pas gepromoveerd? Laat het weten aan onze redacteur.

Redacteur: Nicolaos Starreveld
FNWI, Universiteit van Amsterdam
Postbus 94214
1090 GE Amsterdam
verdediging@nieuwarchief.nl

But, p-values have a downside, they perform very poorly when time is involved. This happens, for example, when studies are accumulated one after the other and decisions are made in between about the timing of new studies and meta-analysis.

Consider the following scenario, a team is carrying out a trial where they want to study whether a beta-blocker prevents a second deadly heart attack. They randomly divide a very large group of patients into two groups of equal size, one that will take the placebo, and one that will take the beta-blocker. The proportion of heart attacks in the placebo group could serve as an indicator of the effectiveness of the beta-blocker. If it has a positive effect we expect a significantly higher proportion of heart attacks in the placebo group, since heart attacks in the beta-blocker group are prevented. But how can you decide whether an observed higher proportion of heart attacks in the placebo group is more likely due to an effective drug, or just occurred by chance?

The p-value tells us something about how unusual such a higher proportion of heart attacks would be in the placebo group, assuming that the beta-blocker is not effective. If we wait for 50, 100, 150, or 200 heart attacks to occur, we expect half of the heart attacks to occur in the placebo group if the beta-blocker is not effective. Due to random fluctuations, however, this can also be a bit smaller or a bit higher. Each possible sample proportion has a probability to occur, and together all possibilities form the sampling distribution shown in Figure 1.

Sampling distributions depend on the sample size n, in this example the number of heart attacks the team of researchers observed. To observe a sampling proportion in $[0.60, 0.62)$ in the placebo group, the probability is different if 50 heart attacks are observed (so 30 of these on placebo, with a probability of 0.042) than if 100 heart attacks are observed (so 60 or 61 of these on placebo, with probability $0.011+0.007$). If a proportion of 0.6 is observed, the p-value is either 0.203, 0.057, 0.018 or 0.006 for 50, 100, 150 or 200 total heart attacks, respectively. The p-value can be computed only if the sample distribution and the sample size are known in advance.

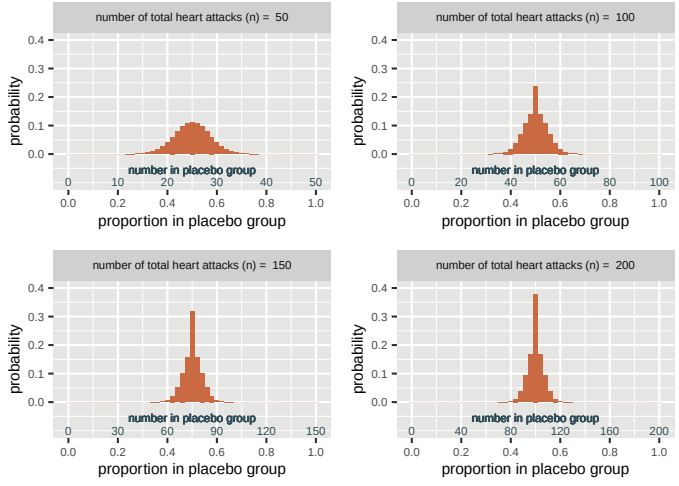


Figure 1 Sampling distributions of the proportion of heart attacks in the placebo group, for various total number of heart attacks (n), under the assumption that the treatment is ineffective (the attacks occur at random in the two groups).

Suppose the team of researchers did a trial, and out of 150 heart attacks they observed a proportion of 0.6 in the placebo group. In that case, the drug seems to prevent heart attacks and the p-value would be 0.018. The beta-blocker looks promising and that might be a good reason to start a new trial. In a follow-up trial, 50 heart attacks are observed, again with a proportion of 0.6 in the placebo group. Combining both studies 120 heart attacks were observed in the placebo group, out of the 200 heart attacks. Can the researchers now state that the p-value is equal to 0.006, as we stated before about sample sizes of 200 total heart attacks? As it would be in a trial where they analyzed all the heart attacks together? The answer is "no"! Because the success of the first study influenced the existence of the second study! This means that only in some selected scenario do we reach two studies and perform the meta-analysis, introducing bias in the sampling distribution. This bias was first described in Judith's work and called *accumulation bias*.

p-value-based statistical tests are intended to be prospective and require the sample size — or the stopping rule that produces the sample — to be specified in advance of observing any data. In the beta-blocker trial, for example, the researchers should predetermine that the study will stop when exactly 200 heart attacks will be observed, also if the first 150 might show a harmfull effect of the drug! Moreover, the use of p-value tests suggests that the results of earlier studies should be unknown when planning new studies as well as planning meta-analyses. Such assumptions are unrealistic. Because most meta-analyses are retrospective and give a statistical summary of the results obtained in all previous studies — after a systematic review has been performed. At the same time, ignoring these assumptions invalidates conventional p-value tests and inflates type-I errors.

ALL-IN meta-analysis resolves the time deficiency in the standard p-value by replacing it with an *e-value*. An e-value is defined as the outcome of a nonnegative random variable with expectation 1 under the null hypothesis. For example, suppose a player is betting on the roulette table, and suppose that the player starts with 1€. If a bet is placed on either black or red, for simplicity forget the zero in roulette, then the expected return after one round is $0.5 \times 2€ + 0.5 \times 0 = 1€$. Hence this betting score in roulette is an e-value.

Accumulation bias and ALL-IN

Accumulation bias is caused by the dependence of follow-up studies on the results of previous studies. Judith showed that all forms of accumulation bias are related to the time aspect in meta-analysis. She developed a framework to describe that accumulation bias is not necessarily a problem, like with standard p-value analysis, but part of the solution!

Judith describes the general form of a test statistic that can withstand any accumulation bias process: the likelihood ratio (which is an e-value). The likelihood ratio offers the meta-analyst the flexibility to decide at any time to finalize the meta-analysis and advise against future studies. It also fosters the accumulation of scientific knowledge, which is critical in reducing research waste.

The heart of an ALL-IN analysis lies in constructing a nonnegative *martingale* (a sequence of random variables for which the conditional expectation of the next observation, given all the previous observations, is equal to the most recent observation)

using likelihood ratios. Each new study adds a term to this martingale process. The likelihood ratio becomes smaller when the observations are more likely under the null hypothesis. For likelihood ratios, using Ville's inequality a threshold can be established that guarantees type-I error control under any accumulation bias process and at any time, as follows:

$$P_0\left(\mathrm{LR}^{(t)} \geq \frac{1}{a} \text{ for some } t = 1, 2, \ldots\right) \leq a,$$

where the likelihood ratios $\mathrm{LR}^{(1)}, \mathrm{LR}^{(2)}, \ldots, \mathrm{LR}^{(t)}$ yield the desired martingale, and the discrete time units correspond to the times at which we add a new study to the meta-analysis.

Communication

Next to her research, Judith emphasizes the need for simple and clear communication of statistical results. As she states in her dissertation: "p-values are turning science into a sorting machine for single studies. Science needs more spirit of collaboration, more efficiency, and simpler communication."

She likes using gambling and betting scores to communicate statistics. Let's see an example concerning the efficacy of vaccines against Covid-19, that illustrates how betting scores can contribute in this direction. More detailed explanations and computations, can be found in Judith's thesis [pp. 20–24].

On 30 June 2020, the US FDA published its guidance document on 'Development and Licensure of Vaccines to Prevent Covid-19'. This set the goals for any Phase III clinical trial on a protective effect of a vaccine against Covid-19. The document prescribed two things to achieve: (1) at least a vaccine efficacy (VE) of 50% and (2) evidence against a null hypothesis of $<30\%$ VE (i.e. the lower endpoint of the corresponding confidence interval should be $>30\%$). According to the FDA, the goal is not only to rule out an ineffective vaccine but also reject the hypothesis that the vaccine has an effect that is too small.

Judith interpreted the design and the results of the Covid-19 vaccine trials using betting scores. Her goal is to bet on one of the two possible outcomes: either the next infection is in the vaccinated or in the placebo group. Using the betting score we can decide whether the vaccine is a real deal-breaker (the scores behave like the salary of a professional poker player) or whether it is not effective enough (the scores behave like anyone playing the roulette wheel). To ensure that the betting scores can show either case, she designed a game that is fair — under the null hypothesis of 30% VE or smaller — and then optimize playing the game with a strategy that is profitable — under the alternative of 50% VE or larger.

Let's see what the betting scores would say in two actual trials, the trial of Pfizer/BioNTech and the CureVac AG mRNA vaccine. The Pfizer/BioNTech trial observed 8 cases of Covid-19 among vaccinated participants, and 162 among the participants assigned to placebo. In the CureVac AG trial, the numbers were $83:145$ vaccinated : placebo. If these results are interpreted using the betting game above, Pfizer/BioNTech could report a total betting score of 118 million €, while CureVac AG has a betting score of 1.84 €, both starting with 1 €. If two players win such amounts at the poker table, who would you consider a professional player with a favorable strategy, and who is the beginner just lucky to be still in the game like anyone playing roulette?

The more personal aspect

As a final note we would like to give the word to the doctorate.

Judith, how did you get interested in statistics?

"From elementary school, I knew I wanted to become a scientist — I used to say 'inventor' then — when I would grow up. My interest in statistics is best described by a quote from statistician John Tuckey: 'The best thing about being a statistician is that you get to play in everyone's backyard.' Statistics is everywhere, and many important decisions are made relying on it, which puts a lot of responsibility on doing it right. I also find it very attractive that statistics can be applied to a broad range of problems. It may be funny but I decided to start this particular PhD project on an evening in a café, where my promoter was giving a talk. I realized there and then that this is the topic I could work on and stay motivated for four years. It has also a societal impact, which I found important, and so it happened that I still enjoy the line of research very much."

Were you also involved in some activities you would like to share with the readers?

"Early 2020, I initiated the ALL-IN-META-BCG-CORONA research initiative. BCG researchers from the university medical centers of Utrecht and Nijmegen were among the first to announce their clinical trial. The main goal was to find whether an immune response to the BCG vaccine, originally developed to protect against tuberculosis, provides indirect protection against Covid-19. The researchers shared their protocol when other researchers around the world started similar trials. Because of the urgency, multiple trials were addressing the same question simultaneously, and a live meta-analysis could provide huge benefits. We contacted the research group and proposed to use our methodology to analyze all these BCG trials together continuously, while they were still ongoing. In this project, the involved trials put collaboration before their interests, since protocols and results were shared before being published. I think that such a collaborative attitude is needed in science, it can increase value, and reduce research waste. All trials are completed now. We hope to put out the results this month (December 2022)."

Concluding

To summarize, in her research Judith developed an ALL-IN meta-analysis, a statistical methodology that makes it possible to analyze results coming from different ongoing studies in a line of research that is growing. Accumulation bias is not a problem but part of the solution, encapsulated in the likelihood ratio and its generalization, the e-value. It also offers the meta-analyst the flexibility to decide at any time to complete the meta-analysis and advise against future studies. Moreover, Judith used betting scores to communicate statistical results simply, clearly, and convincingly.

Judith is currently working as a consultant in biostatistics at Amsterdam UMC where she also continues her research on ALL-IN meta-analysis. She is currently collaborating with researchers in the field of oncology to set up a meta-analysis of onging clinical trials with ALL-IN methodology. We wish Judith all the best with her work on guaranteeing that statistical errors stay under control and results are communicated clearly!

Judith's thesis can be read here: https://ir.cwi.nl/pub/31587.