Introduction

> If I have seen further it is by standing on the shoulders of Giants.
>
> –Sir Isaac Newton, 1675

In the cumulative science we idolize – "standing on the shoulders of giants" – the standard *p*-value is impossible to calculate correctly unless we do clinical trials and meta-analyses for random reasons. This deficiency can be resolved by *ALL-IN meta-analysis*, for *Anytime*, *Live* and *Leading INterim* meta-analysis. Instead of forcing clinical trials into a random walk, this new approach to meta-analysis keeps a *Live* account of what we already know and lets the results so far, even *INterim* results, be the *Leading* source of information on where to go with new studies. This is possible because an ALL-IN analysis deals much better with *time* than any *p*-value analysis can. Together with sequential betting – going all-in, but not all-or-nothing – *time* is the main theme of my Ph.D. research. I will first discuss what I mean by that in the first pages of this dissertation (3-9). I then return to the *p*-values and gambling (pages 9-12), and introduce the contents of the main chapters of this work (pages 12-15).

Time in randomized clinical trials and meta-analysis

Time moves forward; it has a chronology from earlier to later and with more time we increase how much we can observe. If we are learning, we should be able to know more now than we did yesterday. We can consider learning in science as such a process, but with more-or-less discrete units: over time more studies are performed and published. The scientific ideal is that those studies accumulate knowledge and that science itself is cumulative. Moreover, in clinical trial research, the ideal of *Evidence Based Research* (Lund et al., 2016) is that we can also get the *timing* right. We should not passively wait for enough studies to arise to inform medical guidelines in evidence-based medicine, but let those existing studies actively steer the decisions on new research. Sometimes, the time is right to do more studies. Sometimes the time is right to just give a final overview, and declare the line of research completed. Which is which needs to be an evidence-based decision that is informed by a systematic review of all results so far.

Randomized clinical trials *Time* can also be a much more concrete aspect of a scientific study if it simply means it takes time to wait for your observations. Throughout this Ph.D. dissertation, I will consider randomized controlled clinical trials (RCTs) that study two groups of randomly allocated participants. In RCTs, this waiting can be very pronounced. If you study whether a vaccine prevents Covid-19 infections, you have to first vaccinate large groups of participants – half with the vaccine, half with placebo – and then wait for them to get infected. If you study whether vitamins protect against cancer, you have to first convince a large group of participants to add supplements to their diet – half vitamins, half sugar pills – and then wait for cancer. If you study whether a beta-blocker prevents a second deadly heart attack, you have to first start cardiovascular patients on treatment – half on the real drug, half on placebo – and then wait for the deaths. That last one is a classic case in which a systematic review and meta-analysis proved their use, chronicled by Richard Peto in his 1987 address "Why do we need systematic overviews of randomized trials?" (Peto, 1987).

Systematic review and meta-analysis When Richard Peto studied beta-blockers in the early eighties, many trials had sought heart attack patients and followed them for years. Hardly any one of them, however, observed enough deaths to convincingly declare that beta-blockers were protective. Fortunately, Peto and colleagues were able to collect all the trials of sufficient quality and combine their observations in a single analysis: a systematic review and meta-analysis (Yusuf et al., 1985). What defines a systematic review is that it aims to construct a complete collection of the results of all publications that try to answer a similar question. The next step is to evaluate them based on quality such that your selection gives a good impression of what is known so far. A meta-analysis adds to that by giving a statistical summary of the results with a notion of uncertainty, usually in terms of a standard error, confidence interval, or p-value.

P-values over *time*

This is not the place for a full historic account of every peculiarity and misunderstanding that was ever pointed out about the p-value. That would also be an impossible task, and I congratulate Van Dongen and Van Grootel (2021) for writing a comprehensive overview – 70 pages and much more supporting documentation – of the arguments made between 2011 and 2018 in the psychology and psychological methods literature alone. Here, I simply add a point to that discussion: the standard p-value deals very poorly with the aspects of time – chronology and timing – that are so important to cumulative science and evidence-based research.

Before the beta-blocker trials started, each was designed to wait for deadly heart attacks to occur in a beta-blocker treatment group and a control group. If any heart attacks were to be prevented by the drug, this design expects that the proportion of deaths in the placebo group will be larger than that in the beta-blocker group.

So the proportion of heart attacks in the placebo group could serve as a summary of the results[2]. While such a proportion makes sense no matter how many heart attacks we observe, this is not the case for the p-value. Long before we have the data to calculate the proportion, the procedure for the p-value already needs to know the timing of our analysis: how many heart attacks we are going to observe before we calculate the p-value. If that sample size is not fixed in advance or otherwise completely unrelated to (i.e. statistically independent of) the results, the standard p-value makes no sense.

The p-value is a notion of surprise; it tells us something about how unusual our result – our proportion of heart attacks in the placebo group – would be if our treatment is nothing better than a placebo. What is the probability to observe a heart attack in the placebo group in that case? Well, if our treatment is nothing more than a placebo itself, that heart attack can happen to anyone, and since we divided the participants at random over the two groups, also the group will be random. So 0.5 a chance it will be the beta-blocker group, and 0.5 a chance it will be placebo. If we wait for 50, 100, 150, or 200 heart attacks to occur in this scenario, we expect half of the heart attacks to occur in the placebo group. By random chance, however, this can also be a bit smaller or a bit larger. Each possible sample proportion has a probability to occur, and together all possibilities form the sampling distributions shown in Figure 1.

Sampling distributions depend on the sample size n, which in Figure 1 is the number of heart attacks we observe. To observe a sampling proportion of 0.6 in the placebo group in Figure 1, for example, the probability is different if we observe 50 heart attacks (so 30 of these on placebo, with probability 0.042) than if we observe 100 heart attacks (so 60 of these on placebo, with probability 0.011).

We need to know that sampling distribution (like Figure 1) to calculate a p-value. So we can only calculate this p-value if we know the sample size, the total number of heart attacks. The p-value is the probability of the proportion that we observe, e.g. 0.6, together with the probability of all the proportions that are equally or more extreme. Figure 2 considers everything as more extreme that is larger or equal than 0.6, but also everything equal or smaller than 0.4. Figure 2 gives a two-sided p-value and coloring those tails of the distribution gives the familiar picture. If we observe a proportion 0.6, the p-value is either 0.203, 0.057, 0.018 or 0.006 for 50, 100, 150 or 200 total heart attacks respectively.

Accumulating more studies after the first one

Let us assume that the earliest trial studying beta-blockers was able to observe 150 heart attacks over many years before it published its results. If this first trial would be the final trial, no meta-analysis would ever be performed. What type of results could make this first trial the final one? If any heart attacks are prevented by the drug, we expect the proportion of deaths in the placebo group to be larger than half. So if it is smaller than

[2] To keep this introduction simple, I assume here that the number of participants in the placebo and beta-blocker group is very large and that these group sizes stay equal throughout the trial. Chapter 2 discusses the general case in which the number of participants at risk in both groups can change due to left-truncation, right-censoring and decreasing risk set after events occur in a time-to-event analysis.

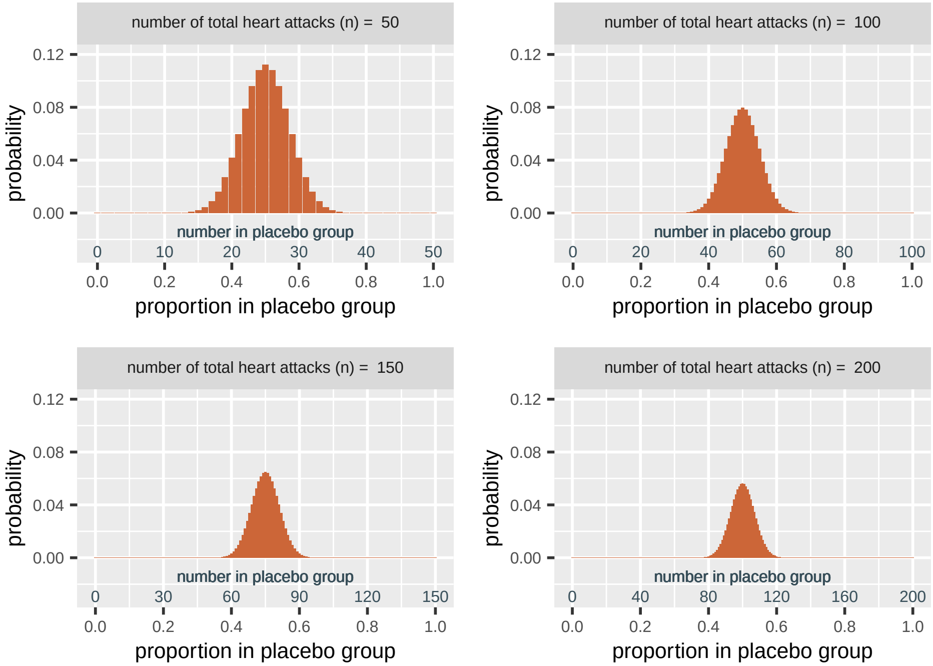

Figure 1. Sampling distributions of the proportion of heart attacks in the placebo group, for various total number of heart attacks (n), if we assume that the treatment is ineffective (the attacks occur at random in the two groups). The number of heart attacks in the placebo group is discrete, such that all possibilities make for 51 possible proportions in the top-left panel ($0/50, 1/50, \ldots, 50/50$) *– each with its probability bar – and 101 possibilities in the top-right panel, 151 in the bottom-left panel and 201 in the bottom-right panel. Because the probability bars add up to a total probability of* 1*, the larger the number of possibilities, the smaller the probability per bar: the height of the bars decreases as their width decreases.*

half, no heart attacks are prevented by the drug, and the drug seems to even cause more heart attacks. If we believe that the drug could be harmful, that is a good reason to not start more clinical trials.

We might want to start more clinical trials if the first study of 150 heart attacks observed a proportion of 0.6 in the placebo group. In that case, the drug seems to prevent heart attacks and the corresponding p-value would be 0.018 as shown in Figure 2. This is usually considered small enough, compared to a level of 0.05 or 5%. So the beta-blocker looks promising and that might be a good enough reason for other researchers to embark on a new trial. But what if the proportion would be smaller than 0.6? Maybe for a proportion smaller than 0.6, nobody would have done a new trial.

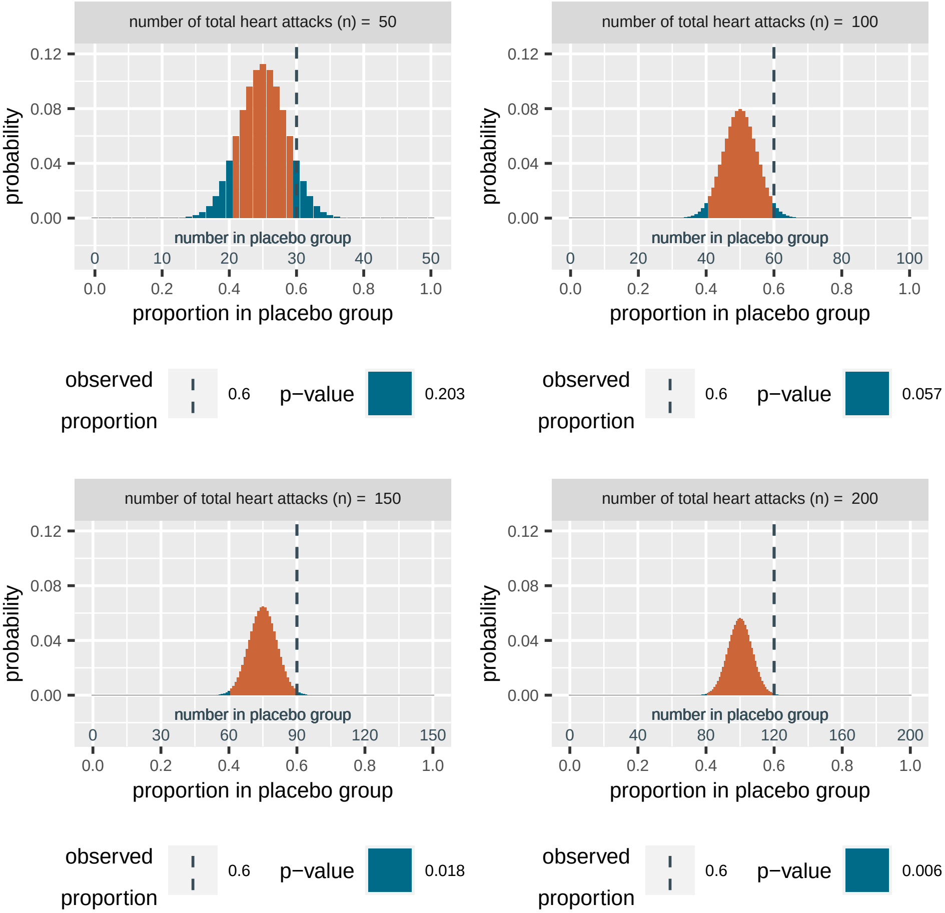

Figure 2. *Two-sided p-value against the null hypothesis of half/half, observing a proportion of heart attacks in the placebo group of 0.6.*

For simplicity, let us assume that these decisions are that clear cut: If the proportion is smaller than 0.6, the drug looks harmful or disappointing, and no more studies are performed. If the proportion is 0.6 or larger, the drug looks promising, and more studies follow.

Meta-analysis timing

A meta-analyst notices that more small studies follow this first study. She decides to systematically collect all these clinical trials, reviews them based on quality, and includes the first trial with a selection of the others in a meta-analysis. In total, these trials observed 200 heart attacks, and coincidentally, out of that total again a proportion of 0.6 occurred in the placebo group. So out of 200 heart attacks, 120 occurred in the placebo group. Can we now calculate our p-value to be 0.006 – like in the right-bottom corner of Figure 2 – if we analyze all the heart attacks together? The answer is "no", because Figure 2 gives the wrong sampling distribution.

There is a process at play that decides whether we even observe the 200 heart attacks that spurred the meta-analysis. The 150 heart attacks in the first study have made the beta-blocker look promising. So more studies follow, but only because the proportion in the placebo group was at least 0.6. The first study would have been the final one if that proportion was smaller than 0.6, in which case we never make it to a meta-analysis of 200 heart attacks. We call such a process an accumulation process and it can happen not only if there is a strict yes/no decision after the early results, but also if disappointing initial results just make it *less probable* for more studies and meta-analyses to follow, instead of completely ruling them out. Whatever the process, it introduces a dependency between the existence and timing of the meta-analysis (e.g. at the total of 200 heart attacks) and the earlier results that are included in that meta-analysis.

Accumulation Bias Such an accumulation process introduces bias into the sampling distribution and can change it completely in comparison to our theoretical distribution. Figure 3 shows what happens in our clear-cut scenario. In this scenario, all the values for proportions smaller than 0.45 (90/200), in the left corner, cannot occur because we already know that we only got to our 200-meta-analysis because the first study showed 90 heart attacks in the placebo group. So all proportions smaller than 0.45 have a probability of 0 and all other possibilities are more probable; the right-hand-side of Figure 3 shows how that shifts the sampling distribution. For a proportion of 0.6, we observed 120 heart attacks in the placebo group and 80 in the beta-blocker group, which is only 30 vs 20 additional ones on top of the first trial with 90 vs 60. The sampling distribution for adding those 50 additional events to the 150 ones we already had is shown in grey. For a meta-analysis on 200 heart attacks, observing a 0.6 corresponds to a p-value of 0.101 for this sampling distribution instead of the p-value of 0.006 from Figure 2.

The p-value is impossible to calculate correctly

What if we are not sure how long we can wait with our analysis; how many heart attacks we will observe before we calculate our p-value? Maybe we first want to observe 50

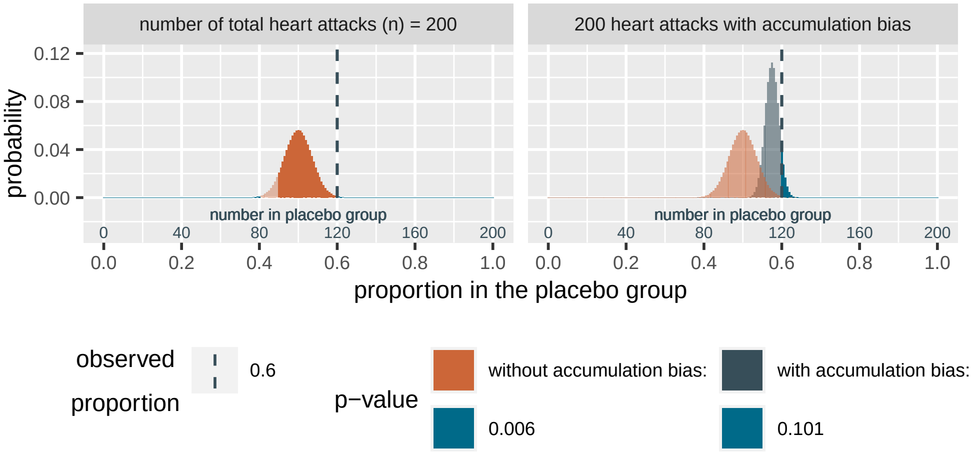

Figure 3. According to the accumulation process we assume, we would not have reached 200 heart attacks in total if we had not already observed 90 in the placebo group in the 150 heart attacks from the first study. So numbers smaller than 90 out of 200 heart attacks in the placebo group are impossible, which is are the proportions smaller than $90/200 = 0.45$.

heart attacks, and then consider if it is worthwhile to extend the experiment and observe another 50. Moreover, as a meta-analyst, we have no control over other researchers that might not know about the results so far and still perform an extra trial. How do I calculate the p-value in that case? Calculating a standard p-value correctly in such scenarios over time is simply impossible. To stick to the familiar calculations from Figure 2 you either have to always stop after $n = 150$ and use that sampling distribution – also if you happen to observe more heart attacks by simply waiting. Or, regardless of the first 150 observations, you continue to $n = 200$ – even if all heart attacks occur in the beta-blocker group and the treatment appears to cause overwhelming harm.

The tricky thing is that we usually do not know – and even cannot know – the grey sampling distribution in Figure 3 or the accumulation bias, the shift in comparison to the theoretical sampling distribution. How much the distribution moves depends on how a research field makes its decisions. The more selective they are – only embark on new studies when the treatment looks promising – the worse it is. But as long as nobody controls how these decisions are made, the standard p-value is impossible to calculate correctly.

Gambling

In gambling, we know that certain things should have a small probability to happen, otherwise the system breaks. If it would be highly probable to win an income at the roulette table for months in a row, casinos could not offer roulette table betting to their customers while staying in business. That does not mean that it is impossible to win an income at the roulette table for months in a row. It could happen, with a very small probability. This is the intuition that this Ph.D. dissertation conveys by expressing statistical evidence in terms of gambling winnings or betting scores.

Time in a casino

In a casino, the *time* it takes customers to play and the *timing* they choose to cash-out their chips are not that important. A casino knows it is not probable to win an income at their roulette table, without specifying how anyone is going to try. A gambler can visit the casino one day a month and another one seven days in a row. Someone else can visit only on Saturdays if the last Saturday was profitable, or the opposite and stay in the casino until they make a profit or bankrupt themselves. For none of these customers is it probable to win an income at the roulette table. If they do, the casino is surprised. The larger their winnings the more surprised the casino is – if their winnings are large in comparison to the initial money they put on the table.

Just like for p-values, we can connect a notion of surprise to a probability statement on how often surprises happen. No matter who is playing at the roulette table, we can predict what the chances are to reach €1000, €100, or €20 in winnings. If we set the goal at €1000, at most 1 in 1000 players will make it if they start with €1. If we set it at €100, at most 1 in 100 makes it when starting with €1. And if we set the goal to €20, at most 1 in 20 will do so. It is that simple (Crane and Shafer, 2020). These statements hold no matter the timing of the good fortune or the betting schedule.

Controlling probabilities

1 in 20 is 0.05 and this 5% reminds us of what we use p-values for: controlling how often something happens that is not supposed to be probable. Standard p-values cannot do this well when *time* is involved; when we accumulate studies one after the other and make decisions in between about the timing of new studies and meta-analysis. In our roulette analogy, however, such decision-making does not matter. We do not need to know how many rounds we played to get to our results, or – even worse: the counterfactual – whether we would have observed the additional round if the results would have been worse earlier on. We can judge gamblers by their winnings and it does not matter how they make their decisions over time and when they plan to quit. Should this matter in a meta-analysis of clinical trials?

ALL-IN meta-analysis and *e*-values ALL-IN meta-analysis resolves the time deficiency in the standard p-value by replacing it with an e-value, a statistic calculated from the data that behaves like the winnings in a casino if we should be surprised to win a lot (like at the roulette table). If we are analyzing beta-blocker trials, we are betting against the probability that the heart attacks occur in the treatment and placebo groups at random. Consistently winning at the roulette table shows that there is something off with the random casino model that makes the ball land on red and black. Likewise in analyzing beta-blocker trials, consistently increasing our e-values shows that there is something off with the random allocation model that makes the heart attacks occur in treatment or placebo. Just like our winnings in the casino, an e-value is our notion of surprise, and it is valid at any time.

Statistical communication: p-value by picture

In the introduction to Chapter 1, I disappoint my friends and former self (see Preface) by stating we are actually back to proposing p-values. Specifically, our winnings in betting and *e-values* – can be interpreted as conservative p-values by taking their inverse.

In Appendix Section 1.A I clarify this by stating that we are not talking about the standard type of p-value, as it is presented in introductory statistics texts and can be intuitively pictured as in Figure 2. I believe this is important for the field working on betting scores and e-values: distinguish between the standard (strict) introductory-textbook p-value-by-picture and the more general, abstract, mathematical definition to be found in advanced and mathematical statistics texts[3] – especially if the abstract definition allows for p-values with properties that you should not expect from the introductory-textbook p-value-by-picture.

A statistics audience might want to know how a new concept relates to existing ones. The general audience, however, only vaguely remembers what p-values even are. A vague memory assigns attributes that p-values do not have, like being a posterior probability for the null hypothesis, or being the probability of a type-I error. It is better to remind the audience of p-value-by-picture than to generalize it. The picture forces us to think about the sampling distribution before we can do the tail-coloring. This sampling distribution reminds us that we need a sample size (or stopping rule) in advance. This is a very important limitation of standard p-values.

Even though based on generalizing the abstract definition even further, it is possible to formulate anytime-valid p-values (Johari et al., 2021), I believe we should focus elsewhere with so many anti-p-value feelings around. Especially in a meta-analysis, where we judge a line of research instead of a single publication, p-values are more of an enemy than a friend. The perils of meta-analysis lie in publication bias and selection bias, and p-values are the main tool for the underlying questionable research practices of file-drawing and p-hacking. P-values were never designed to do so, but they are turning science into a sorting machine for single studies. Science needs more spirit of collaboration, more efficiency, and simpler communication of the evidence so far and what more is necessary. Standard p-values are not so helpful in this regard. I believe that betting scores and e-values are and that they can stand on their own.

Is gambling a good idea?

Professor Glenn Shafer's work on game-theoretic inference served as a major inspiration for my Ph.D. research. We both think that, in communicating statistics, thinking about gambling helps with intuitions about uncertainty in a way that is somehow natural, or

[3] The mathematical definition allows for conservativeness (the probability that p is smaller than α can be much smaller than α) and does not require a fixed sample size.

"part of our cultural upbringing", even if you do not gamble yourself. Just like Glenn (SIPTA, 2021), I have never been a gambler, and not even entered a casino at any point in my life. As a statistician, I definitely do not play the lottery, so with *intuition for* I do not mean *believe in*.

There are quite some similarities between running a beta-blocker clinical trial and playing poker. For one thing, you might feel a need to convince the outside world that what you are doing is worthwhile. Some people disapprove of deciding a patient's treatment by a coin toss, but would not recognize routine treatments can be just as uncertain to benefit as to harm them (Evans et al., 2011). Some people disapprove of making your salary in casino-located poker tournaments, but would not recognize that we take risks in about every major life decision (Konnikova, 2020). Both clinical trials and poker teach us things about decision making that are is important enough to write books about, with titles as *The Biggest Bluff: How I Learned to Pay Attention, Master Myself, and Win*, *Thinking in Bets: Making Smarter Decisions When You Don't Have All the Facts* and *The Signal and the Noise: Why So Many Predictions Fail – but some don't*. Poker is a form of gambling in which it is possible to play a strategy that turns the odds in your favor. And you can also play it just to learn how to make better decisions in the future.

Contents of this Ph.D. dissertation

ALL-IN meta-analysis stands for *Anytime*, *Live* and *Leading INterim* meta-analysis. ALL-IN 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, *live* – no need to prespecify the looks, and *leading* – in the decisions on whether individual studies should be initiated, stopped or expanded, the meta-analysis can be the leading source of information.

Going ALL-IN

The phrase *going all-in* comes from poker and means that we move – or *shove* or *jam* – all our chips onto the table and risk them in the round of the game we are playing. Going all-in can be necessary to force other players into folding when they cannot match our bet (*call*) or raise it. Professional poker players use this move while knowing they have a savings account full of backup money – a bankroll. So the all-in move is part of a strategy that will not bankrupt them in the long run over many tournaments but is aggressive enough to get an edge if they play it well. This long-run of tournaments is what makes professional poker different from an all-or-nothing game in which losing would mean that you can never play again.

In poker, you need to practice and pay attention as the game progresses and the observed moves accumulate (Konnikova, 2020). As you get better, you get richer, and you can turn to tournaments with a larger buy-in. So as your knowledge accumulates, your money accumulates. In fact, how much you win is a good proxy for how well you play. This is not the case in games of pure chance, like the lottery or roulette. It is this distinction that drives ALL-IN meta-analysis.

Chapter 1 ALL-IN meta-analysis

This introductory chapter presents ALL-IN meta-analysis to a broad audience interested in statistics. It formulates the null and alternative model involved in statistical testing, defines a precise gambling game, and shows that we can formulate betting scores that behave like casino betting under the null hypothesis of no treatment effect in clinical trials. It generalizes the statistics from discrete observations – such as heart attacks and infections – to general meta-analysis methodology that is based on Z-score approximations. Apart from statistical testing, it also introduces anytime-valid analysis for estimation with confidence intervals.

This first chapter was very much influenced by the Covid-19 pandemic that showed that science is a major gamble. If we do not accept that and do not play a coordinated strategy, we end up with enormous research waste. Examples are the hundreds of clinical trials on hydroxychloroquine in ICU patients (Glasziou et al., 2020) and the long wait – while passing the mark of 2 million deaths worldwide – for published results on other treatments, like budesonide (Yu et al., 2021). ALL-IN meta-analysis allows improving a future pandemic response as well as non-pandemic evidence-based medicine in terms of statistics, efficiency, collaboration, and communication.

Chapter 2 The Safe logrank test

This chapter goes deeper into the machinery of betting scores and e-values that make ALL-IN meta-analysis possible for trials that observe events like heart attacks, tumor recurrence, and Covid-19 infections, i.e. time-to-event analysis. It shows that we can construct an e-value logrank test under the assumption of proportional hazards and that these ideas can be extended to confidence intervals for the hazard ratio and meta-analysis based on summary statistics.

This chapter is more technical and contains the necessary derivations to show that the e-values we propose can construct test martingales and be used for anytime-valid statistical analysis. However, the chapter also contains many figures that compare the rejection regions and sample size needed to those of existing approaches to do sequential logrank testing. Using a Gaussian approximation on the logrank statistic, we illustrate that the safe logrank test (which itself is always exact) has the same type of rejection region to O'Brien-Fleming α-spending but with the potential to achieve 100% power by optional continuation. Although the approach to *study design* requires a larger sample size, the *expected* sample size is competitive by optional stopping.

Chapter 3 Accumulation Bias

This chapter returns to accumulation bias and describes this phenomenon in its generality in an accumulation bias framework. This allows us to model a wide variety of practically occurring dependencies, including study series accumulation, meta-analysis timing, and approaches to multiple testing in living systematic reviews. The strength of this framework is that it shows how all dependencies similarly affect p-value-based tests. Accumulation Bias in meta-analysis is inevitable, and even if it can be approximated and accounted for, no valid p-value tests can be constructed.

This chapter also shows that the problem of accumulation bias is not new. To some extent, it has been recognized in the clinical trial literature, but not confronted. The accumulation bias framework helps to recognize the approaches to handle accumulation bias. While e-values and ALL-IN meta-analysis provide one way, by considering error control that stays intact – "survives" – over time as we add more studies, another way is to use priors and do a Bayesian analysis and condition on results and the timing itself.

Chapter 4 and Chapter 5

These two chapters were written as blog posts and explain the two approaches to handle accumulation bias with examples, simulation R code, and figures. These chapters introduce an extreme version of accumulation bias that allows for simpler notation and easy simulation. As such, these chapters can be read as a more accessible presentation of the ideas in Chapter 3.

These two chapters have the same introduction as they discuss exactly the same example accumulation bias process, but with two different ways of counteracting it. Chapter 4 gives more detail about what it means for a scientific field to handle accumulation bias using ALL-IN meta-analysis, which Chapter 3 presents as error control *surviving over time*. Chapter 5 gives more detail about what it means for a scientific field to handle accumulation bias by Bayesian analysis, which Chapter 3 presents as error control *conditioned on time*. Both can use ALL-IN e-values. The first as a notion of evidence that has type-I error control averaged over all sizes of study series. The second as a notion of evidence that has Bayesian error control conditioned on the study series, by using it as a pseudo-Bayes factor combined with prior odds. The specification of prior odds does make this second approach more difficult in a retrospective meta-analysis, and we discuss the risks when information in the data seeps into the prior odds. The appendix to Chapter 5 contains the proof for using e-values as pseudo-Bayes factors for Bayesian error control, which is a new technical result that we would like to expand in a paper in the future.

Chapter 6 Data sharing in a live meta-analysis

This chapter details my experience of performing an ALL-IN meta-analysis on actual clinical trials during the Covid-19 pandemic. It discusses the practical constraints of running a *live* meta-analysis in terms of data sharing while ensuring privacy and blinding.

This ALL-IN meta-analysis studied whether the Bacillus Calmette-Guérin (BCG) vaccine, originally developed to protect against tuberculosis, could also protect against (severe) Covid-19 infections. It started with a clinical trial in The Netherlands that was soon replicated in many countries around the world. Because these trial investigators around the world were in close contact, I could propose to run a live analysis in a large collaboration. The trial statistician from Utrecht University Medical Center, dr. Henri van Werkhoven, became the meta-analysis Principle Investigator, with dr. Alexander Ly from CWI and myself as the meta-analysis statisticians. We formed a steering committee with prof. dr. Marc Bonten (Utrecht UMC), prof. dr. Mihai Netea (Radboud UMC, Nijmegen) and prof. dr. Peter Grünwald and all trials participated in regular Advisory Committee meetings.

The collaboration started in the Spring of 2020 and is still ongoing. Results of the analysis will be published later this year, so this chapter only details operational considerations, not the data. This meta-analysis did not produce press-attention-grabbing recommendations early in the pandemic, but – maybe because of that – can still be considered a scientific success. It is an example of evidence-based research since all individual trials were involved in evaluating the body of research over time, so automatically placing their results in the context of the evidence base. This improved the value of the studies and prevented research waste.

Discussion and future work

The discussion section relates the ideas in this Ph.D. dissertation to statistical standards at Cochrane, the leading authority on meta-analysis of clinical trials. It considers reasons why sequential meta-analysis was discussed thoroughly, but never implemented, and how updating meta-analyses can lead to decisions on "redundancy" of future clinical trials – an example of research waste. The discussion concludes with a reflection on possible future research.