ALL-IN
meta-analysis

Judith ter Schure

ALL-IN meta-analysis

Judith ter Schure

ISBN 978-90-619-6413-1

Cover design

Ilse Modder

Cover photo credits

Marjolein van Sommeren (conceptualization), Kilian Lafleur (digital edit),
Rinske ter Schure (photographer), Thomas de Jong (photographer),
Elsa ter Schure (photographer), Arnold ter Schure (encouragement)

ALL-IN meta-analysis

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op donderdag 7 april 2022
klokke 15:00 uur

door

Julia Anna (Judith) ter Schure
geboren te Meppel, Nederland
in 1992

Promotores:

Prof. dr. Peter D. Grünwald (Universiteit Leiden en
Centrum Wiskunde & Informatica, Amsterdam)

Dr. Daniel Lakens (Technische Universiteit Eindhoven)

Promotiecommissie:

Prof. dr. Frans A.J. de Haas

Prof. dr. Jelle J. Goeman

Dr. ir. Joanna in 't Hout (Radboud Universiteit)

Prof. dr. Glenn Shafer (Rutgers University)

Prof. dr. Alex J. Sutton (University of Leicester)

This work was funded by the Dutch Research Counsil (NWO) and carried out at
Centrum Wiskunde & Informatica (CWI), Amsterdam.

Centrum Wiskunde & Informatica


SD564683
50
20
10

To Glenn Shafer, Stephen Senn, Peter Grünwald and Daniel Lakens

Subsets of you taught me the importance of fundamentals and history, the beauty of clinical trials, and – especially when a mathematical concept is necessary to make a point – the power of storytelling.

Origin of the material

The dissertation is based on the following earlier (pre-print) publications:

Chapter 1 is based on a paper that is under review at F1000 and available on ArXiv:

> Judith ter Schure and Peter Grünwald. ALL-IN Meta-analysis: Breathing Life into Living Systematic Reviews. arXiv:2109.12141. 2021.

Chapter 2 is based on a paper that is available on ArXiv:

> Judith ter Schure, Muriel F. Pérez-Ortiz, Alexander Ly and Peter Grünwald. The Safe Logrank Test: Error Control under Continuous Monitoring with Unlimited Horizon. arXiv:2011.06931. 2020.

Chapter 3 is based on a paper that is published at F1000 Research:

> Judith ter Schure and Peter Grünwald. Accumulation Bias in Meta-analysis: The Need to Consider Time in Error Control [version 1; peer review: 2 approved]. *F1000Research*, 2019.

Chapter 4 is based on a blogpost at The Replication Network:

> Judith ter Schure. Accumulation Bias: How to Handle It ALL-IN. *The Replication Network*. 2020.

Chapter 5 is based on a blogpost at The Replication Network:

> Judith ter Schure and Peter Grünwald. Accumulation Bias: How to Handle It As a Bayesian. *The Replication Network*. 2022.

Chapter 6 is based on a paper published in STAtOR, the society magazine of the Netherlands Society for Statistics and Operations Research VVSOR:

> Judith ter Schure, Peter Grünwald and Alexander Ly. Pandemic Preparedness in Data Sharing: Lessons Learned from Collaborating in a Live Meta-Analysis. *STAtOR*, 2021, 22.4: 47-52.

Contents

Preface

This Ph.D. research had its origin in a bar; a typical bar in Utrecht, in a historic wharf cellar at the central canal. On Wednesday, April 20th 2016, this bar served as the scenery for the Young Statisticians to host their night of beers and statistical discussion on the (ab)use of p-values in research: "*To p or not to p?*" It was there that I heard Professor Peter Grünwald speak about how p-values are misunderstood and how much better we could do if we thought of statistics a bit more like gambling. I enjoyed every minute of it – also thanks to the great atmosphere that evening – and, fortunately, I still do.

Later that year I finished my Master's *Statistical Science for the Life and Behavioural Science* while staying in contact with Peter. I was very lucky that the timing of my graduation matched with Peter's procurement of funding for Ph.D. students. As a contender for a position, I had the advantage to have already made my job interview impression that day in that bar. Peter remembers it as quite unorthodox in mathematics for a student to simply walk up to him and state something along the lines of "This is so cool! Can I spend a Ph.D. studying this?".

Now, almost four years of Ph.D. research[1] later, I am still not bored with p-value discussions. What is more, friends refer to my Ph.D. research as "the nemesis of the p-value", and they have a point. What else could be the final blow to "science by p-value" than a paper (Ter Schure and Grünwald (2019), Chapter 3) that points out that in the cumulative science we idolize – "standing on the shoulders of giants" – the p-value is impossible to calculate correctly unless we do clinical trials and meta-analyses for random reasons?

[1] Four full-time equivalent years: between May 1st, 2017 and February 1th, 2022 I spent 44 months working 80% of my working week ($\approx$ 35 weeks full-time equivalent) on this Ph.D. research and 13 months working 100%, so 48 full-time months in total.


5
BCE ECB EUS EZB EKP EKT ESB EKB BCE EBC 2017
10

Introduction

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

In the cumulative science we justly admire, 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 (5-11). I then return to the *p*-values and gambling (pages 11-14), and introduce the contents of the main chapters of this work (pages 14-17).

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.

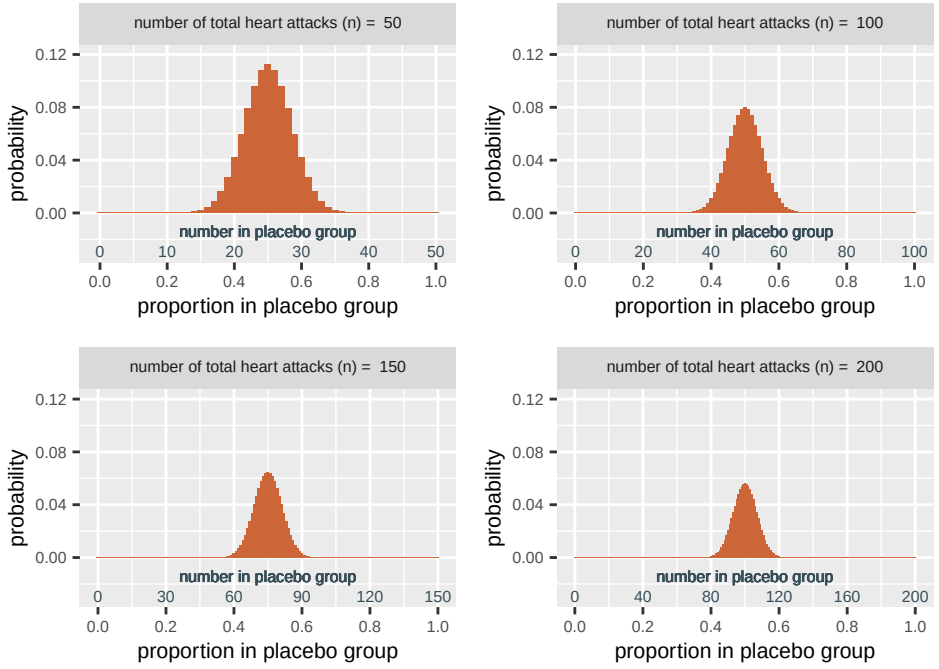

Figure 1. Sampling distributions of the proportion of heart attacks in the placebo group, for various total number of heart attacks (n), if we assume that the treatment is ineffective (the attacks occur at random in the two groups). The number of heart attacks in the placebo group is discrete, such that 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, like 0.55? Maybe for a proportion smaller than 0.6, nobody would have done a new trial.

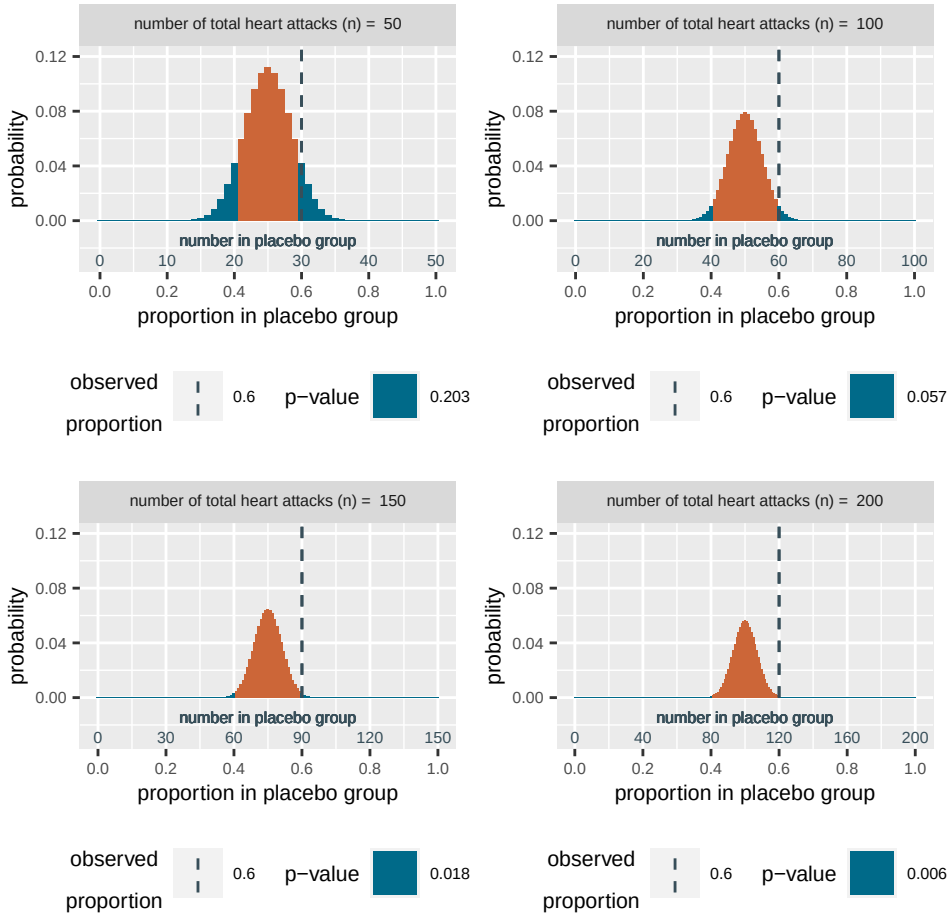

Figure 2. *Two-sided p-value against the null hypothesis of half/half, observing a proportion of heart attacks in the placebo group of 0.6.*

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

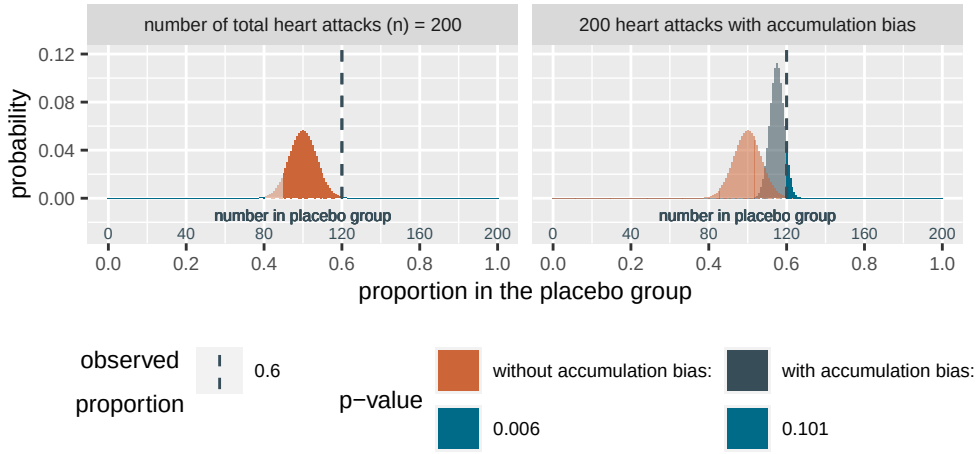

Figure 3. *According to the accumulation process we assume, we would not have reached 200 heart attacks in total if we had not already observed 90 in the placebo group in the 150 heart attacks from the first study. So numbers smaller than 90 out of 200 heart attacks in the placebo group are impossible, 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 on page 1) 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" – James Bond, the Oceon's movies, The Hangover – even if you do not gamble yourself (there's even gambling in Toy Story). 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 share 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 size of 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 pseudo-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. This chapter 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.

1 | ALL-IN meta-analysis

Abstract

Science is justly admired as a cumulative process ("standing on the shoulders of giants"), yet scientific knowledge is typically built on a patchwork of research contributions without much coordination. This lack of efficiency has specifically been addressed in clinical research by recommendations for living systematic reviews and against research waste. We propose to further those recommendations with *ALL-IN meta-analysis*: *Anytime Live* and *Leading INterim meta-analysis*. ALL-IN provides 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. We illustrate the method for time-to-event data, showing how synthesizing data at *interim* stages of studies can increase efficiency when studies are slow in themselves to provide the necessary number of events for completion. The meta-analysis can be performed on interim data, but does not have to. The analysis design requires no information about the number of patients in trials or the number of trials eventually included. So it can breathe life into living systematic reviews, through better and simpler statistics, efficiency, collaboration and communication.

Introduction

The scientific response to the Covid-19 pandemic constitutes a major gamble. In the US, for example, the funding program for vaccine development did not put money on a single vaccine, but on six different ones. They purposely took "multiple shots on goal" according to Larry Corey of the NIH Covid-19 Prevention Network in an interview with STAT (Branswell, 2021). Vaccine development is not a sure thing, and so their strategy needed to be robust enough to just "let the chips fall". Also in the search for treatments, the scientific community had to hedge its bets. Clinical trials competed for resources and patients, and had to continuously change course when new information arrived. In contrast to vaccines, however, in most countries a strategy to find treatments was lacking. Many clinical trials suffered from "poor questions, poor study design, inefficiency of regulation and con-

duct, and non or poor reporting of results": research waste (Glasziou et al., 2020). We believe that more strategic thinking can benefit a future pandemic response as well as non-pandemic evidence-based medicine, as uncertainty is often a given. Honest scientific bets can breathe life into the approach called *living systematic reviews* that aims to keep the evidence record up-to-date (Elliott et al., 2017) and the medical guidelines current (Akl et al., 2017). We propose to make those bets by using *ALL-IN meta-analysis* in clinical trial design, monitoring and reporting.

ALL-IN meta-analysis stands for *Anytime Live* and *Leading INterim* meta-analysis. The *Anytime* aspect provides analysis that controls type-I error in testing and coverage in interval estimation regardless of the decision making along the way, and so regardless of any stopping rules or accumulation bias processes (Chapter 3). The *Live* aspect prevents research waste caused by meta-analyses that are out-of-date, which is often the case in retrospective meta-analysis. The synthesis can be a bottom-up collaboration of trials, as well as a prospective top-down statistical analysis for decision making. The *Leading* aspect allows the systematically collected evidence included so far to drive the necessity and design of new trials. Finally, the *INterim* aspect is new in meta-analysis and makes for effortless combination of trials while they are still ongoing. What is more, ALL-IN meta-analysis is also literally *ALL-IN* since any number of new studies can be included; it has an unlimited horizon. We illustrate this in the setting of time-to-event data, where waiting for new events is an inherent challenge of clinical trials. Combining trials early can prevent delays if studies are slow in themselves to complete the necessary number of events. ALL-IN has advantages in four categories: statistics, efficiency, collaboration and communication. We introduce all four briefly (page 23-25) before we go into more detail, but first illustrate the language of betting for single trials studying a Covid-19 vaccine.

A single trial: the FDA game

On June 30th, 2020, the US Food and Drug Administration (FDA) published its guidance document on "Development and Licensure of Vaccines to Prevent Covid-19" (FDA, 2020). This set the goals for any Phase III clinical trial betting on a protective effect of a vaccine against Covid-19. The guidance document advised on the definition of events of confirmed (symptomatic) SARS-CoV-2 infection for the trials to be counting. And in counting those, the document prescribed the two things to achieve: (1) at least a vaccine efficacy (VE) of 50% and (2) evidence against a null hypothesis of $\leq$ 30% VE (FDA, 2020, p. 14). Most Covid-19 vaccine trials randomized large numbers of participants 50:50 vaccine:placebo such that we can assume that also throughout the trial the participants at risk stayed approximately balanced. According to the definition of SARS-Cov-2 infections, we start counting once a participant has a confirmed infection after being fully vaccinated for at at least a number of days (e.g. 7 days in the Pfizer-BioNTech trial (Polack et al., 2020))). This is also when a (virtual) bet could start. In the following we reinterpret the design for the Covid-19 vaccine trials in the language of betting.

Each new event carries evidence that we express by a betting score. We make a (virtual) investment on one of the two outcomes: either the next event occurs in the vaccine group or it occurs in the placebo group. If there is no effect of the vaccine whatsoever, the 50:50

risk set ensures that the infected participant has 0.5 a chance to be vaccinated and 0.5 a chance to be a placebo. Yet, following the FDA, we do not only want to rule out an ineffective vaccine, but also reject the hypothesis that the vaccine has an effect that is too small – set as the null hypothesis of (at most) 30% VE. In that case each newly observed infection has slightly smaller chance to be a vaccinated participant. That probability to be in the vaccine group is 0.41, since each placebo group member has a 100% risk of Covid-19 and a vaccine group member has $100-30=70\%$ of the risk, which is a fraction 0.41 of the total risk ($70/(100+70)$). So if the VE is too small to be of interest we expect (at least) a fraction 0.41 of Covid-19 events to occur in the vaccine group and (at most) 0.59 in placebo.

How do we bet against that and win if the vaccine has a much larger protective effect? We are betting *against* the probability 0.41 of the next Covid-19 event to occur in the vaccine group. If this probability actually is that large (the vaccine is not very protective; the null hypothesis) we do not want the game to be favorable under any strategy, just like the casino does not want any gambler to earn a salary playing the roulette wheel. On the other hand, we are betting *in favor* of a much smaller probability for the vaccine group. If this probability is smaller (the vaccine is protective; the alternative hypothesis) we do want to win money, just like a professional poker player who makes a salary out of gambling well. We use the betting scores to 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 our betting scores can show either case, we first *design* the game such that it is fair – under the null hypothesis – and then *optimize playing* the game with a strategy that is profitable – under the alternative.

Designing a fair game under the null hypothesis Consider gambling at the roulette table where the vaccine trial analogy is like betting on red (vaccine) or black (placebo). Betting correctly doubles your investment, betting incorrectly loses everything you risked. Assuming no house edge (no 0 or 00 on the roulette wheel, on which you cannot bet) and an initial €100 you do not expect to increase your investment, since you have 0.5 a chance of doubling ($2\cdot$€100) and 0.5 a chance of losing all ($0\cdot$€100). Whether you bet everything on black or red, in expectation the betting score after one round is $(0.5\cdot 2+0.5\cdot 0)\cdot$€100, which is the initial investment €100. To achieve the same thing betting against the 0.41:0.59 probabilities instead of 0.5:0.5, your investment needs to multiply by 2.4 (1/0.41) for vaccine and 1.7 (1/0.59) for placebo. If you bet everything on vaccine you have 0.41 chance of multiplying by 2.4 ($2.4\cdot$€100) and 0.59 chance of losing all ($0\cdot$€100) and if you bet everything on placebo you have 0.59 chance of multiplying by 1.7 ($1.7\cdot$€100) and 0.41 chance of losing all ($0\cdot$€100). The expected betting score after one round is again the initial investment for both: $(0.41\cdot 1/0.41+0.59\cdot 0)\cdot$€100 and $(0.59\cdot 1/0.59+0.41\cdot 0)\cdot$€100. Hence, at either the roulette table or in this FDA game, by design the game is fair and does not favor us. After all, if our observed infections land on the vaccine and control group with the probabilities 0.41:0.59 – like a spin of the roulette wheel on black and red with 0.5:0.5 – we do not expect to claim an effective vaccine.

Optimize playing the game under the alternative hypothesis How do we win as fast and as much as possible if our observed infections do not behave like a roulette wheel? It has been known since the work of Kelly (1956) and Breiman (1961) that the best way to increase your capital in the long run is to not bet all your (virtual) investment €100 on one of the two possible outcomes (red/vaccine or black/placebo) but to divide it based on the odds that make the game favorable to you. So our focus needs to be on the minimal VE of 50% from the FDA guidance. In the scenario of 50% VE, the probability that the next Covid-19 case is in the vaccine group is 1/3: if we set the risk of Covid-19 for a placebo group member to 100%, a vaccine group member has $100-50=50\%$ of that risk, which is 1/3 of the total risk ($50/(100+50)$). Kelly (1956) and Breiman (1961) urge us to invest one-third ($1/3\cdot$€100) on observing the next infection in the vaccine group and two-thirds ($2/3\cdot$€100) on placebo.

Likelihood ratios If we bet this way we can rewrite our betting scores in terms of a *likelihood ratio*. We first show this for the red-black roulette game where we double what we had put at risk on either black or red if the spin of the roulette wheel outputs the color we bet on. Just like in our strategy in the FDA game, we put $1/3\cdot$€100 on red and $2/3\cdot$€100 on black, so we win the following if the ball X lands on either **red** or **black**:

$$X=\textbf{red} \qquad 2\cdot\frac{1}{3}\cdot€100=\frac{\mathfrak{L}(1/3\mid X)}{\mathfrak{L}(1/2\mid X)}\cdot€100$$

$$X=\textbf{black} \qquad 2\cdot\frac{2}{3}\cdot€100=\frac{\mathfrak{L}(1/3\mid X)}{\mathfrak{L}(1/2\mid X)}\cdot€100$$

The Bernoulli 1/3-likelihood $\mathfrak{L}(1/3\mid X)$ assigns likelihood 1/3 when is $X=$ **red** and 2/3 when is $X=$ **black**. So if our strategy is to invest 1/3-2/3 in roulette, our payout is our initial investment €100 multiplied by the likelihood ratio, whether X is **red** or **black**.

$$X=\textbf{vaccinated} \qquad 2.4\cdot\frac{1}{3}\cdot€100=\frac{\mathfrak{L}(50\%\text{ VE}\mid X)}{\mathfrak{L}(30\%\text{ VE}\mid X)}\cdot€100$$

$$X=\textbf{placebo} \qquad 1.7\cdot\frac{2}{3}\cdot€100=\frac{\mathfrak{L}(50\%\text{ VE}\mid X)}{\mathfrak{L}(30\%\text{ VE}\mid X)}\cdot€100$$

The likelihood for 50% VE ($\mathfrak{L}(50\%\text{ VE}\mid X)$) assigns likelihood 1/3 when is $X=$ **vaccine** and 2/3 when is $X=$ **placebo**. Similarly, the likelihood for 30% VE ($\mathfrak{L}(30\%\text{ VE}\mid X)$) assigns likelihood 0.41 when is $X=$ **vaccine** and 0.59 when is $X=$ **placebo**. Hence if our strategy is to invest 1/3-2/3 in the FDA game, our payout is also our initial investment €100 multiplied by the likelihood ratio, whether X is **vaccine** or **placebo**.

A winner We assume now that we start with an initial (virtual) investment of €1 instead of €100. At the first observation we bet €0.33 on vaccine and €0.66 on placebo. After we observe the event in the placebo group we lose our €0.33 bet on vaccine and multiply our €0.66 on placebo by 1.7 to €1.13. The likelihood ratio between our 30% VE alternative hypothesis and our 50% VE null hypothesis – so $\mathfrak{L}(50\%\text{ VE}\mid X)/\mathfrak{L}(30\%\text{ VE}\mid X)$ – is also about 1.13, so multiplying our initial investment of €1 into €1.13. On the other hand, if

we observe the event in the vaccine group we lose our €0.66 bet on a placebo event and multiply our €0.33 on vaccine by 2.4 to €0.81. The likelihood ratio of a vaccine event multiplies our investment by 0.81. After each observed event we reinvest what we have left in the new bet, so multiply that with the next likelihood ratio.

The Pfizer/BioNTech trial observed 8 cases of Covid-19 among participants assigned to receive the vaccine and 162 cases among those assigned to placebo (Polack et al., 2020), so they could report a total betting score of $0.81^8 \cdot 1.13^{162} \cdot$ €1, which is about €118 million (note that 1.13 is really 1.13333...)[1]. If someone wins that at the poker table, we have good reason to consider her a professional poker player with a favorable strategy, rather than a lucky beginner (Konnikova, 2020).

Meta-analysis: bottom-up collaboration

The Pfizer/BioNTech trial included more than 43 thousand participants (Polack et al., 2020), which is quite unique for a clinical trial. Usually trials are much smaller, and scientific consensus is built through systematic reviews and retrospectively combining trials in a meta-analysis. ALL-IN is a way to do so by collaborating bottom-up in a strategic way that can be live instead of retrospective. It has advantages in four categories that we will first briefly introduce and then further elaborate on in this chapter: statistics, efficiency, collaboration and communication.

Statistics

Not all mRNA vaccines showed such favourable results as the Pfizer/BioNTech vaccine. In a press release CureVac AG (2021) announced that the final analysis of their clinical trial observed 83 events in the vaccinated group and 145 in placebo, so only a 43% VE[2] (our calculations assuming a 50:50 balanced risk set (r = 1 in CureVac AG (2020, p. 124))). Their protocol states the FDA goal in terms of a confidence interval that excludes a VE of 30%, adjusted for two interim analyses. That adjusted confidence interval at the final analysis is based on $Z_{\alpha/2}$-statistic for the nominal level $\alpha/2 = 0.02281$ (CureVac AG, 2020, Table 8). That interval is [25.3%, 57.1% VE] (our calculations; normal approximation interval) and, regrettably, does not exclude 30%. When the chips fell, this trial lost.

Statistical analyses like these are essentially *all-or-nothing*, just as any other $p < \alpha$ analysis. As soon as all the α is spent – either on a few interims and a final analysis or just on one fixed sample size – we cannot continue the trial and perform subsequent analyses without violating the type-I error rate. This might be a reasonable price to pay in the urgency of a pandemic when multiple vaccines are competing, but it is a very inconvenient property for clinical trials in general. Usually, we do want to reanalyze a clinical

[1] We can ask: why did the Pfizer/BioNTech trial not declare efficacy sooner? In a later agreement with the FDA in October, the vaccine trials committed to collecting at least two months of safety data for half of the included participants in the trial. So the trial could not stop earlier. This agreement was formalized in the FDA guidance document for Emergency Use Authorization and is still present in the May 2021 version (FDA, 2021)

[2] The CureVac AG (2021) press release reports a VE of 48%, so uses a different r (ratio of follow-up time in the two groups). In such large trials r can often be assumed to stay close to 1, so we set it to 1 to make all calculations simpler. All our calculations can be found via Appendix Section 1.B.

trial in combination with other similar trials in a meta-analysis. Yet any $p < \alpha$ procedure is equivalent to setting a rejection region for the test statistic and checking whether the value for the statistic falls within that region. This rejection region is based on a sampling distribution that assumes the number of studies in the meta-analysis, and the number of patients within each study to be fixed in advance. Given such a fixed sample size (but also for any sequential stopping rule that sets a maximum sample size in advance, such as α-spending), there is only one region, and your test statistic is either in it or not. If it is not, you are not allowed to redo the analyses with an increased sample size. This problem is recognized in approaches to control type-I error for living systematic reviews (Simmonds et al., 2017). But also if the meta-analysis is not updated, the α is essentially already spent on the individual trial, since the meta-analysis is an update of the trial analysis that is unscheduled and lacks type-I error control at the same level α. If the individual study analysis would have been conclusive, the meta-analysis might never be performed, and we can recognize that we are dealing with a situation of "meta-optional-stopping". A different way to see this is by the actual sampling distribution of trials in a meta-analysis: any data-driven decision within the series – whether to accumulate more studies and when to perform the meta-analysis – changes the sampling distribution and invalidates the fixed-sample-size distribution assumed for $p < \alpha$. Hence hardly any meta-analysis has valid type-I error control, when the accumulation of trials is based on strategic decisions that introduce accumulation bias (Chapter 3).

ALL-IN meta-analysis is not *all-or-nothing* but can still combine all available studies. In fact, it allows any number of new studies or patients to be included without ever spending all α. In terms of gambling, we can keep betting our virtual investment because we never lose everything. The CureVac AG (2021) results, for example, would have accumulated a betting score of $0.81^{83} \cdot 1.13^{145} \cdot$ €1 = €1.84. This single trial is not very profitable, but at least it still preserves some evidence to reinvest in the next trial, such that we can continue to observe evidence and express it by betting on additional observations in a new trial. An ALL-IN meta-analysis can always continue testing the null hypothesis – with type-I error control – and estimating the confidence interval – with coverage guarantees. Importantly, for these tests and intervals the procedures are exactly the same no matter what decisions – so-called stopping rules, or accumulation bias processes (Chapter 3) – are at play.

Efficiency

Lack of efficiency has been addressed in clinical research in many ways. Not only in the proposal of living systematic reviews (Elliott et al., 2017), but also in encouragements to present new studies in the context of existing evidence (Young and Horton, 2005), in advice to design new trials based on systematic reviews and meta-analysis (Chalmers and Lau, 1993; Lau et al., 1995; Sutton et al., 2007; Goudie et al., 2010; Lund et al., 2016) and in pleading to prevent the "scandal" of wasteful research into clinical questions that are already answered or not of primary importance (Altman, 1994; Chalmers and Glasziou, 2009; Glasziou and Chalmers, 2018; Glasziou et al., 2020, "research waste"). These calls have not been completely ignored, since clinical research has seen an increase in efficiency – e.g. in platform trials or adaptive meta-analysis (Tierney et al., 2021) – when-

ever collaboration is deemed possible prospectively. Nevertheless, most clinical trial data is synthesized retrospectively, and still deserves all of the above recommendations. ALL-IN meta-analysis enables these data-driven decisions that can make science more efficient. New studies can be easily informed by the synthesis of all data so far such that exactly the right number of patients are randomized to answer a research question, no more and no less. Moreover, an ALL-IN meta-analysis can give an account of the evidence at anytime and therefore facilitate prioritizing new studies, if more than one line of research needs additional data, but not all can be funded.

Collaboration

ALL-IN meta-analysis can be a *live* meta-analysis, since it does not matter how many studies will eventually be combined or which study will contribute most data. Whether it is based on summary statistics (Tierney et al., 2021) or on individual patient data (IPD) (Polanin and Williams, 2016), involvement in the same meta-analysis facilitates discussion between those running trials in the same line of research; especially if the line of research can be concluded early. Trial protocols and statistical analysis plans can be exchanged and scrutinized, to identify discrepancies between the design and the conduct of trials. In an ongoing meta-analysis, trials can be selected for inclusion before investigators are unblinded to the results, which helps to mitigate the problems of publication bias and p-hacking. If IPD analysis is possible, intense collaboration might also prevent mistakes and fraudulent data that would otherwise depreciate the meta-analysis.

A meta-analysis benefits from homogeneity. With too much heterogeneity, it can be very disheartening to update a random-effects meta-analysis, since many trials are needed to precisely estimate the between trial variation and overcome it (Sutton et al., 2007; Kulinskaya and Wood, 2014; Jackson and Turner, 2017). Close collaboration might prevent unnecessary heterogeneity, if trial investigators are involved in the selection of trials in the meta-analysis; especially if they can advise on the design and conduct of new trials and align inclusion criteria and endpoint definitions. A fixed-effects meta-analysis can conclude the research effort early. Sufficient homogeneity may be possible in close collaboration.

Communication

The language of betting The interpretation of evidence in terms of a betting score might help to communicate the uncertainty in statistical results. As Shafer (2021) puts it: "When statistical tests and conclusions are framed as bets, everyone understands their limitations. Great success in betting against probabilities may be the best evidence we can have that the probabilities are wrong, but everyone understands that such success may be mere luck." Thinking in terms of bets also helps to understand when statistical analyses can be *anytime-valid*. If they are of the *all-or-nothing* kind, but reanalyzed in a meta-analysis, they are gambling while broke. (This intuition can be made mathematically precise; see the description of Neyman-Pearson testing in terms of betting Shafer (2021) and Grünwald et al. (2019).) Yet if we add new studies to an ALL-IN meta-analysis, we are reinvesting the betting score that we saved from earlier studies, to evaluate whether the strategy in

those earlier studies continues to succeed. Just like when reinvesting your profits in a casino from one slot machine into another, the notion of winning stays the same. Our evidence against the hypothesis of a *fair* casino does not change when we alternate slot machines. It does not change if we use the score so far to decide on alternating them or to decide when to cash out. If the slot machines are fair, any strategy of playing them is not expected to make money, and our notion of type-I error control holds under any dependency on past results (stopping rules or accumulation bias processes).

Other communication Those uncomfortable with the language of betting can also easily resort to any of three more familiar notions of statistical communication. Firstly, the likelihood ratios/betting scores and their generalizations, so-called *e-values* (Grünwald et al., 2019; Vovk and Wang, 2021), can be interpreted as conservative p-values by taking their inverse. If we denote any betting score or e-value by $€$ (e.g. $€ = €1.84$ for the CureVac trial data), then $p < 1/€$ is a conservative p-value (e.g. $p = 1/1.84 = 0.54$ for the CureVac trial data). If we communicate the p-value $p = 1/€$ anyone can test by comparing $p < \alpha$ but with the addition that this conservative p-value is anytime valid[3] and so $p < \alpha$ can never spend all α (it is never an *all-or-nothing* test). Secondly, the likelihood ratios have their own notion of evidence in the likelihood paradigm (Royall, 1997). Just as well as stating that the Pfizer/BioNTech trial (Polack et al., 2020) multiplied €1 to almost €118 million and the CureVac AG (2021) trial multiplied €1 to €1.84, we can state that their data was almost 118 million times and 1.84 times more likely if we assume the FDA's goal of 50% VE in comparison to assuming only 30% VE. For Pfizer, that sounds very good, for CureVac, not so much, and so these numbers have an interpretation of their own without imposing any α level. Thirdly, likelihood ratios can be accepted by the Bayesian paradigm, as Bayes factors, and possibly combined with prior odds. Grünwald et al. (2019) and Grünwald (2021) show that betting scores/e-values and Bayes factors are closely related, although not all Bayes factors are betting scores/e-values. The bottom-line for communication purposes is that the reporting by ALL-IN meta-analysis can be interpreted in many ways – p-values, likelihood ratios, Bayes factors – but regardless of the interpretation provide fully frequentist type-I error control for tests and coverage for confidence sequences.

The remainder of this chapter discusses the four categories of advantages in more detail: *Statistics* in Section 1.1, *Efficiency* in Section 1.2, *Collaboration* in Section 1.3 and *Communication* in Section 1.4. We use the Covid-19 vaccine trials as running examples, based on the FDA game described already, but also in terms of the *safe/e-value logrank test* (Chapter 2). We also briefly discuss an actual ALL-IN meta-analysis in Section 1.3.1, that used this safe/e-value logrank test to study whether the BCG vaccine could protect against Covid-19 (Van Werkhoven et al., 2021, ALL-IN-META-BCG-CORONA). In the concluding section we will provide some broader context, with an overview of all the methods already developed – e-values, *safe* tests (Grünwald et al., 2019) and *anytime-valid* confidence sequences – methods already available in software – notably safestats R package

[3] Such conservative p-values cannot be pictured as the tails of a sampling distribution since such a picture needs a sample size. Appendix Section 1.A gives more details.

(Turner et al., 2022) – and future work. R code for all calculations, simulations and plots is online available via Appendix Section 1.B.

1.1 Statistics

The language of betting comes with the intuition that winning a large betting score a has small probability if the null hypothesis is generating our observations (e.g. the roulette wheel is fair). We will make this intuition precise and show how to control the type-I error by bounding this probability by Markov's inequality and Ville's inequality. Crucial here is that the betting score underlying our test is an *e-value*. The language of betting also comes with the intuition that when playing a game that is favorable to us in principle, we can use strategies of different quality: even among all strategies under which we expect to get richer, some of them can be expected to earn us much more than others. We will relate the more well known notion of power to such a different notion of *optimality*. In the following we discuss both *e-values* and *optimality* first for a single trial (in the FDA game and more generally) and then for ALL-IN meta-analysis. We conclude by a generalization of optimal e-value tests to confidence sequences.

1.1.1 Under the null: e-values in a single trial

To make the FDA game fair we imposed a multiplication by 2.4 (or 170/70) if we observe the event in the vaccine group and 1.7 (or 170/100) if we observe it in the placebo group. This multiplication has expectation 1 (or smaller) if we assume the null hypothesis of a vaccine with negligible VE of 30% (or smaller). In case of 30%, we have probability 0.41 (or 70/170) to observe a vaccine event and probability 0.59 (or 100/170) to observe placebo, so in expectation we multiply our investment by: $70/170 \cdot 170/70 + 100/170 \cdot 170/100 = 1$. No matter how we invest in the two outcomes, (e.g. putting 1/3 on vaccine and 2/3 on placebo, or something different) in expectation under the null we multiply the initial investment by 1. This means that our betting score is an e-value, since by definition an e-value is the outcome of a nonnegative random variable with expectation 1 under the null hypothesis (Grünwald et al., 2019).

Our betting score could also be rewritten as a likelihood ratio, so the expectation of the likelihood ratio $(\mathfrak{L}(50\%\text{ VE} \mid X)/\mathfrak{L}(30\%\text{ VE} \mid X))$ is 1 as well. We henceforth write the likelihood ratio after n rounds of betting (or after observing n events) as $\mathbf{LR}^{(n)}$, with for the FDA game

$$\mathbf{LR}^{(n)} = \prod_{i=1}^{n} \frac{\mathfrak{L}(50\%\text{ VE} \mid X_i)}{\mathfrak{L}(30\%\text{ VE} \mid X_i)}. \tag{1.1}$$

Using its expectation of 1, Markov's inequality bounds the probability of observing a large multiplication of our investment € (a large likelihood ratio) by α after $n = 170$ rounds as follows:

$$\mathbf{P}_{30\%\text{ VE}}\left[\mathbf{LR}^{(170)} \geq 1/\alpha\right] \leq \frac{\mathbf{E}_{30\%\text{ VE}}\left[\mathbf{LR}^{(170)}\right]}{1/\alpha} = \frac{1}{1/\alpha} = \alpha.$$

Figure 1.1 shows at the right side the histogram of betting scores in the FDA game after 170 events when we simulate events under the null hypothesis, with probability 0.41 to

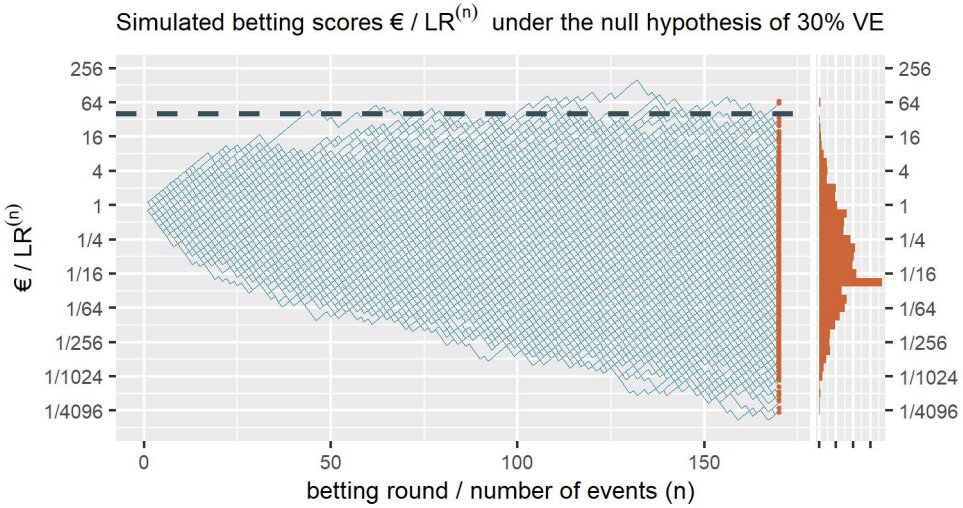

Figure 1.1. 1000 simulated betting scores in the FDA game over betting rounds n assuming a probability of 0.41 *(*70/170*) for each event to occur in the vaccine group (the null hypothesis of* 30% *VE). The dashed line is the threshold* $1/\alpha = 40$ *one-sided. The histogram at the right shows the betting score/*$LR^{(170)}$ *after 170 events. Note that the expectation of* 1 *of the scores is not the mode of its distribution nor its median and that the y-axis is on a log scale.*

occur in the vaccine group, corresponding to 30% VE. A line is shown at 40, and indeed no more than $\alpha = 1/40 = 2.5\%$ of the scores seem to be larger than that threshold. In fact, in these 1000 runs of simulation only 0.3% of the runs have a betting score larger than 40; Markov's inequality is a loose bound. We also have a stronger result because we obtained our betting score over events by multiplying the score of the rounds (see (1.1), corresponding to reinvesting our winnings), called Ville's inequality. We get the following from Ville (1939):

$$\mathbf{P}_{30\%\,\mathrm{VE}}\left[\mathbf{LR}^{(n)} \geq 1/\alpha \quad \text{for some } n\right] \leq \alpha.$$

Ville's inequality is also illustrated in Figure 1.1: if we take the sequence of rounds into account, still only a few out of the 1000 simulations *ever* reach a betting score larger than 40. In fact, in these 1000 runs of simulation only 1.1% of the runs have a betting score that is larger at any round in the game, such that our type-I error is controlled at $\alpha = 2.5\%$ at any time. Moreover, this type-I error control is not tied to this maximum number of 170 events, but continues to hold with an unlimited horizon. Making a large profit in such a fair game casts doubt on the null hypothesis and is captured by a likelihood ratio that grows away from 1: a large betting profit is obtained if the null likelihood is performing worse than alternative.

When trials can be summarized as bets Before they can be combined in a meta-analysis, individual trials are often characterized by the summary statistics from trial pub-

lications. Conventional meta-analysis combines these statistics (e.g. mean differences and standard deviations) in a Z-statistic (Borenstein et al., 2009). Unlike the vaccine/placebo outcomes that we have seen so far, such a Z-statistic has a continuous density and cannot be summarized by separately dealing with all possible outcomes. Fortunately, Shafer (2021) shows that any likelihood ratio of distributions can be viewed as a betting score in a game with initial investment €1. This is possible because likelihood ratios have expectation 1 in general if we assume the null hypothesis in the denominator of the ratio. For a Z-statistic we have two normal distributions with variance 1, one with mean μ_0 under the null hypothesis, and one with μ_1 under the alternative. If the data is generated by the null model, the expectation of the likelihood ratio is

$$\mathbf{E}_{Z\sim\phi_{\mu_0}}\left[\frac{\phi_{\mu_1}(Z)}{\phi_{\mu_0}(Z)}\right] = \int_z \phi_{\mu_0}(z)\frac{\phi_{\mu_1}(z)}{\phi_{\mu_0}(z)}\mathrm{d}z = \int_z \phi_{\mu_1}(z)\mathrm{d}z = 1, \tag{1.2}$$

since $\phi_{\mu_1}(z)$ is a probability density that integrates to 1. This means that any such likelihood ratio for a Z-statistic is an e-value and can be used to construct tests by betting.

Not all summary statistics can be assumed to form a Z-statistic with a normal distribution. Fortunately for the logrank statistic this is reasonable (Chapter 2) if studies are large and the effect size not too extreme (hazard ratios not too far away from 1). We will use the logrank Z-statistic as a running example for meta-analysis on summary statistics. For an IPD meta-analysis (on individual patient data), however, we recommend to use the exact safe/e-value logrank test from Chapter 2 that is valid regardless of the randomization (e.g. 1:1 balanced or 1:2 unbalanced), the number of participants at risk, the number of events or the size of the effect – so also for a hazard ratio 0.05 that corresponds to a VE of 95%.

1.1.2 Under the null: e-values in a (live) meta-analysis

Assume we want to perform a meta-analysis and we collect a Z-statistic Z_i from each trial i, e.g. a logrank statistic. Before observing Z_i we construct an honest bet $\mathbf{LR}_i = \phi_{\mu_1}(Z_i)/\phi_{\mu_0}(Z_i)$ for each trial that is an e-value and thus has type-I error control under the null hypothesis ϕ_{μ_0} – for a default logrank statistic this is always $\mu_0 = 0$ corresponding to hazard ratio of 1. If we think of the betting score from the first study and invest it in the second study, we are in fact multiplying likelihood ratios. We need to have a notion of time t, such that at each time we know the number of studies $k\langle t\rangle$ so far and the number of observations $n_i\langle t\rangle$ in each study i. If we assume that all studies are completed at time t with $n_1, n_2, \ldots, n_i$ events summarized by logrank Z-statistics $z_1^{(n_1)}, z_2^{(n_2)}, \ldots, z_k^{(n_k)}$ we can construct our ALL-IN bet as follows:

$$\mathbf{LR}_{\mathrm{META}}^{\langle t\rangle} = \prod_{i=1}^{k\langle t\rangle} \mathbf{LR}_i^{(n_i)} = \prod_{i=1}^{k\langle t\rangle} \frac{\phi_{\mu_1\sqrt{n_i}}(z_i^{(n_i)})}{\phi_0(z_i^{(n_i)})}. \tag{1.3}$$

The global null hypothesis Each trial bet is testing the same null hypothesis $\mu_0 = 0$ in (1.3), such that the ALL-IN meta-analysis bet tests a *global null hypothesis* of no effect

(0% VE) in all trials. Such a global null hypothesis can be rejected with a contribution from each trial, but also in case only one trial observes a large score betting against the hypothesis and no other trial observes a very small betting score that loses those winnings again. After all, the null in each trial is rejected as soon as the null is rejected in one of the trials.

Meta-analysis on interim data We can generalize this ALL-IN meta-analysis bet of completed trials to bets on interim data by assuming that we only have an interim logrank Z-statistic $z_1\langle t\rangle, z_2\langle t\rangle, \ldots, z_k\langle t\rangle$ for the $n_1\langle t\rangle, n_2\langle t\rangle, \ldots, n_k\langle t\rangle$ events observed so far at time t; $k\langle t\rangle$ still represents the number of studies so far at time t, but now these studies are not (all) completed. We construct our ALL-IN bet in a similar way:

$$\mathbf{LR}_{\text{META}}^{\langle t\rangle} = \prod_{i=1}^{k\langle t\rangle} \mathbf{LR}_i^{(n_i\langle t\rangle)} = \prod_{i=1}^{k\langle t\rangle} \frac{\phi_{\mu_1\sqrt{n_i\langle t\rangle}}(z_i\langle t\rangle)}{\phi_0(z_i\langle t\rangle)}. \tag{1.4}$$

From the perspective of Ville's inequality, the analysis on completed trials and the one on interim data are indistinguishable. The only thing that matters is that we include all the data we have so far at time t, such that we have type-I error control

$$\mathbf{P}_0\left[\mathbf{LR}_{\text{META}}^{\langle t\rangle} \geq 1/\alpha \quad \text{for some } t\right] \leq \alpha, \tag{1.5}$$

for the global null hypothesis probability $\mathbf{P}_0$ with an unlimited horizon over time t.

1.1.3 Under the alternative: optimality in a single trial

A power analysis sets a very specific goal for a trial, usually to detect an effect of minimal clinical relevance. This is the effect we would not like to miss if it were there, although we hope that the real effect is larger. We nevertheless use this smallest effect of interest to decide on the sample size of the trial, otherwise we risk a futile trial. The FDA was clear on what this minimal effect should be for the Covid-19 vaccine trials: a VE of 50% (FDA, 2020). This is the effect we used to bet in the FDA game.

Our strategy in the FDA game, however, was not trying to achieve optimal power. If we compare the *all-or-nothing* confidence interval for CureVac AG (2021) from the introduction – the final analysis on 83+145 events – we notice that this confidence interval [25.3%, 57.1% VE] is smaller than the final anytime valid interval we show in Figure 1.3 in Section 1.1.5, which is [20.2%, 60.3% VE]. Note that we are comparing a $Z_{\alpha/2}$-confidence interval for $\alpha/2 = 0.02281$ with $\alpha/2 = 0.05$, so the wider interval cannot be attributed to the level of α. The difference is that the former one is optimized to have spent all α at the final analysis, while the latter one is optimized to continue data collection. Power is the probability of finding the desired result using the specified analysis at a sample size or stopping rule. So for an analysis that is intended to have unlimited horizon, power is not a well-defined concept. Instead Grünwald et al. (2019) introduced the concept of *growth-rate optimality in the worst case*, or *GROW*. Here, the goal is to optimize the expected rate at which the evidence grows (or the interval shrinks) for each new data point, not at a specific sample size. The worst case here is the 50% VE for a one-sided alternative hypothesis $H_1 = \{P_{\text{VE}} : 50\% \leq \text{VE} \leq 100\%\}$. We optimized the FDA bet in the introduction

by putting this 50% VE in the alternative likelihood. This can be rewritten in terms of a likelihood ratio for the logrank statistic Z as follows:

$$\mathbf{LR}^{(n)} = \prod_{i=1}^{n} \frac{\mathfrak{L}(50\%\text{ VE} \mid X_i)}{\mathfrak{L}(30\%\text{ VE} \mid X_i)} = \frac{\mathfrak{L}(50\%\text{ VE} \mid X_1, \ldots, X_n)}{\mathfrak{L}(30\%\text{ VE} \mid X_1, \ldots, X_n)} \approx \frac{\phi_{\mu_{\min}\sqrt{n}}(Z^{(n)})}{\phi_{\mu_0\sqrt{n}}(Z^{(n)})}, \tag{1.6}$$

with $\mu_{\min} = \log(0.5)/4$ and $\mu_0 = \log(0.7)/4$ with 0.5 and 0.7 the hazard ratios corresponding to VE of 50% and 30% respectively (see Chapter 2). So our one-sided alternative hypothesis for the logrank Z-statistic is a Z-distribution with a mean representing an effect that is at least $\mu_{\min}$:

$$H_1 = \{\phi_{\mu_1} : \mu_1 \leq \mu_{\min}\}$$

(since positive VE corresponds to a negative μ_1). Our choice of the parameter of the alternative likelihood $\mu_{\min}$ follows directly from the minimal effect set by the FDA. Kelly (1956) already showed that this way of betting optimizes the way our betting score grows if the true VE is 50% (our worst-case scenario). Breiman (1961) showed that this approach also minimizes the number of events we need to reach a given betting score set in advance (e.g. €$1/\alpha$), for which some intuition is given in Figure 1.2. Grünwald et al. (2019), Shafer (2021) and Appendix Section 2.B give various other reasons why this is the best way to bet, relating it to data compression, information theory, Neyman-Pearson testing, Gibb's inequality, and Wald's identity. The most crucial property for the purposes of ALL-IN meta-analysis is that the alternative likelihood puts some money on each possible outcome, such that no matter what outcome we observe, we keep some of the money we risk. This contrasts the approach with a classic $p < \alpha$ test that essentially puts all money on the rejection region, such that if the outcome is not in it, we we lose all and cannot continue betting. A thorough interpretation of Neyman-Pearson testing and p-values in terms of betting is given by both Grünwald et al. (2019) and Shafer (2021).

1.1.4 Under the alternative: optimality in a meta-analysis

ALL-IN meta-analysis allows for a retrospective meta-analysis that is bottom-up. The betting score that we accumulate by reinvesting from one trial into the other (which is multiplying betting scores) has an interpretation without enforcing a common design or stopping rule on all included trials. This is especially important if trials have their own stopping rules, or if meta-accumulation processes are at play that influence the existence of trials based on earlier (trial) results in the same meta-analysis. While a meta-analysis can be bottom-up and each have its own design and effect of minimal interest, it can be advisable to agree on a $\mu_{\min}$ for the meta-analysis. However, the meta-analysis betting score can also allow each trial i to have its own alternative likelihood with parameter $\mu_{\min(i)}$. Then the following multiplication of those betting scores is still a valid meta score with type-I guarantees:

$$\mathbf{LR}^{\langle t \rangle}_{\text{META}} = \prod_{i=1}^{k\langle t \rangle} \frac{\phi_{\mu_{\min(i)}\sqrt{n_i}}(z_i^{(n_i)})}{\phi_0(z_i^{(n_i)})}. \tag{1.7}$$

As long as $\phi_{\mu_{\min(i)}\sqrt{n_i}}$ is a probability density that integrates to 1, we have that each likelihood ratio integrates to 1 under the global null hypothesis, such that (1.5) holds. This

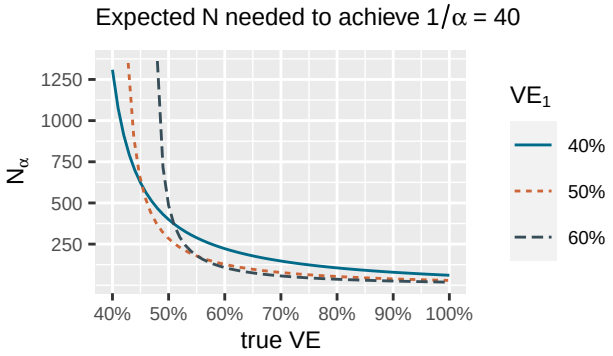

Figure 1.2. *N_α is the expected number of events needed to reach a betting score of $1/\alpha = 40$ for $\alpha = 0.025$ if we bet according to VE_1 indicated by the three different lines, with bets each of the form $\prod_{i=1}^{N} \frac{\mathfrak{L}(VE_1 \mid X_i)}{\mathfrak{L}(30\%\ VE \mid X_i)}$. The number of events we need decreases if the true VE underlying the data increases (the true difference in risk between vaccine and control is larger). The smallest number of events for a true VE of 40% is reached by betting VE_1 of 40% (blue solid line), the smallest number of events for a true VE of 50% by betting VE_1 of 50% (orange dotted line) and the smallest number for true VE of 60% by betting VE_1 of 60% (grey dashed line). Note that for the alternative in the FDA game $H_1 = \{P_{VE} : 50\% \leq VE \leq 100\%\}$ we are only interested in playing the game well if the true VE is 50% or larger. Since for larger true VE, taking $VE_1 = 50\%$ performs quite well, our strategy is to optimize for the worst case of 50% VE itself and use the bet with $VE_1 = 50\%$ in the FDA game.*

means that trials can also learn their parameter $\mu_{\min(i)}$ from already completed trials. This is sometimes the case if trials are not powered to detect an effect of minimal interest, but an effect that is plausibly true based on earlier research. Kulinskaya et al. (2016) shows that such use of existing studies to power new trials can actually bias conventional meta-analysis since it introduces yet another dependency between sample size and results that is unaccounted for in any analysis that assumes a fixed sample size. For ALL-IN meta-analysis this is no problem at all, and trials can learn from each other as long as the parameter $\mu_{\min(i)}$ is fixed before seeing new data that is evaluated using that parameter in (1.7). In Chapter 2 we discuss the advantages of even learning the parameter within one trial using prequential plugins or Bayesian posteriors. In a game like the FDA game with a clear goal, this is inferior to the GROW approach, but in other situations it could be preferred.

1.1.5 Confidence sequences

The CureVac AG (2021) trial reached their final interim analysis but was not able to reject the null hypothesis of 30% VE. The trial had been too optimistic and powered for 60% instead of 50% VE (CureVac AG, 2020). If a trial is underpowered but still has a large number of participants in follow-up, there is good reason to continue the trial, or combine the trial with results from a new trial in a meta-analysis. However, with a total of 227

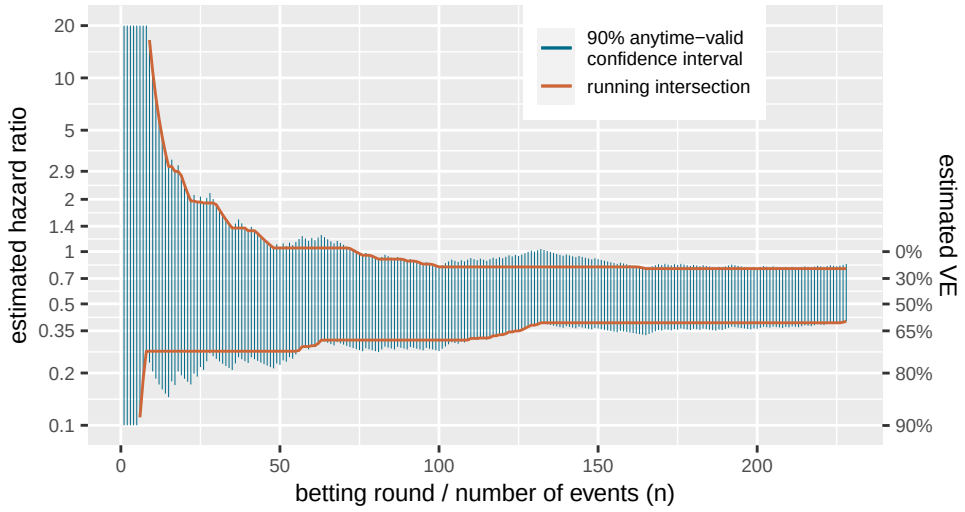

Figure 1.3. *90%-confidence sequence for a random ordering of the* 83 *events in the vaccine group and* 145 *events in placebo from the CureVac AG (2021) trial. Note that the y-axis is on the log scale.*

events this trial was not underpowered to reject the null hypothesis with an effect in the same ballpark as the Pfizer/BioNTech trial that reported 95% VE. In such a case it is very interesting to zoom in on the estimate for the effect, instead of its test.

A standard confidence interval can be seen as an inversion of a hypothesis test: if the null falls outside a two-sided 90%-confidence interval it can be rejected with a one-sided type-I error level of $\alpha/2 = 0.05$. In general, the interval excludes all the values for the parameter that can be rejected. Similarly, in our context, an anytime-valid confidence interval excludes all values of the parameter that can be rejected by the e-value test that corresponds to the betting strategy at hand. So the interval is essentially tracking a whole range of bets, each against a different null hypothesis. Figure 1.3 gives a sequence of anytime-valid confidence intervals for a random ordering of the CureVac AG (2021) data, one for each new observed event or betting round. It shows that the more events we observe, the more parameters (hazard ratios, or their corresponding VEs) we can exclude from the interval. Because these intervals are valid at any time, once we can exclude a value, we never have to include it again. So we also show a sequence of intervals that is the running intersection of all the previous intervals. This of course crucially depends on the ordering, so the one shown for the CureVac AG (2021) data is just an example, since the ordering is not real. Since these intervals are anytime valid, it is possible to further shrink the intervals by continuing follow-up and observing more events. The coverage of an anytime-valid confidence sequence – like an e-value test – has an unlimited horizon.

An ALL-IN meta-analysis confidence interval that is based on a running intersection is of course only possible in an IPD meta-analysis, and cannot be based on summary statistics.

The confidence interval shown in Figure 1.3 is based on the logrank Z-statistic (by repeatedly calculating it after each event), which can also be a summary statistic to achieve a single interval that is anytime-valid. The interval follows from the likelihood ratio of normal densities from (1.6) and follows a general recipe for constructing confidence sequences from Howard et al. (2021) where the hazard ratio is obtained by means of the Peto estimator (Peto, 1987). The same approach can be used to obtain an ALL-IN meta-analysis confidence interval. A fixed-effects meta-analysis Z-statistic corresponds to a logrank statistic stratified by trial, and an estimate can be obtained from such a logrank statistic that Peto (1987) calls a *typical* hazard ratio. We discuss this approach a bit further in the final section.

1.2 Efficiency

Trials often suffer from recruitment difficulties, with estimates of 35% (between 1994 and 2002) and 56% (between 2004 and 2016) not reaching the goal set in advance (McDonald et al., 2006; Walters et al., 2017). These trials find themselves underpowered according to their own protocol: when they decide the stop the recruitment and obtain the final sample size for analysis, they have a high probability for their test statistic to fall outside the rejection region they set in advance. This is exactly the scenario where meta-analysis could rescue the line of research by combining multiple underpowered trials. However, the literature on *research waste* (Chalmers and Glasziou, 2009) and *Evidence-Based Research* (Lund et al., 2016) shows that we are not using the existing evidence base well to design the new trials needed for conclusion or to interpret new research. ALL-IN meta-analysis makes this very easy to do. It comes with a simple notion of the evidence already collected and what is still needed, and a notion of a new trial's ability to provide that: the implied target. The combination of the two has the capacity to make study design more honest, showing what a trial can add to the existing evidence base instead of just evaluating a misguided goal to single-handedly answer a research question.

1.2.1 The evidence so far and what is still needed

An ALL-IN meta-analysis can set a prospective goal for conclusion, e.g. $\alpha = 0.0025 = 0.05^2$ corresponding to the level of α required by authorities like the FDA that ask for two trials at the $\alpha = 0.05$ level. Following Ville's inequality (1.5) we need a betting score of $1/\alpha = €400$ if we start with €1 to reach a conclusion. Because an ALL-IN meta-analysis combines trials by reinvesting or multiplying betting scores, a very simple calculation gives the betting score we still need at any given point. If an initial trial is able to reach a score of €8, any new trial can be designed to multiply that by 50. So on its own, starting with €1 instead of €8, it would need a betting score of €50 to help the meta-analysis reach €400. We could evaluate the sample size of the new trial on its ability to reach 50, which for a fixed sample size gives the conditional power of the meta-analysis once the new trial is added. However, if this second trial also foresees recruitment issues, it is more difficult to evaluate its planned contribution since it will probably not be the final trial in the meta-analysis. For this, Shafer (2021) proposed a new notion for the ability of a test. Rather than stating the probability of reaching a specific target score, we state a sort of

expectation for the target score itself. This is called the *implied target*.

1.2.2 The ability of a new trial: the implied target

The likelihood ratio summarizes the data not in just two categories – statistical significant or not statistical significant – but captures the evidence so far on its way to a certain threshold. Similarly we propose to not evaluate experimental design as all-or-nothing, but summarize its ability to build on what is already there and facilitate future research. To capture a study's expected contribution to a series of studies, we formulate the *implied target* from Shafer (2021) as the multiplicative amount with which the combined evidence is expected to grow if the study – designed with a certain $\mu_{\min}$ and sample size n – is added. In general, the implied target E^* is defined as:

$$E^* = \exp\Big(\mathbf{E}_{Z^{(n)} \sim \phi_{\mu_{\min}\sqrt{n}}}\Big[\log\big(\mathbf{LR}^{(n)}(Z^{(n)})\big)\Big]\Big). \tag{1.8}$$

The logarithm appears in equation (1.8) because the distribution of a likelihood ratio based on n events is very non-symmetric and heavy tailed, with extremely large likelihood ratios occurring with only small probability (see Figure 1.4). So the expectation of the likelihood ratio is drawn very far from its typical values by these large likelihood ratios and is not a good expression of what to expect. The logarithm makes the distribution more symmetric (asymptotically (for large n) and for normal likelihood ratios even normally distributed), such that the expectation is a more meaningful summary of the evidence promised by the study. By exponentiation (exp()) we bring this expectation back to the scale of the likelihood ratio, such that it can be interpreted as a betting score or e-value.

In the FDA game the expected growth rate per new event in the CureVac trial, assuming their effect of minimal interest of 60% VE, is the following:

$$\begin{aligned}&\exp\left(\mathbf{E}_{60\%\text{ VE}}\left[\log\left(\frac{\mathfrak{L}(50\%\text{ VE} \mid X)}{\mathfrak{L}(30\%\text{ VE} \mid X)}\right)\right]\right)\\ &= \exp\left(\frac{40}{140}\cdot\log\left(\frac{50/150}{70/170}\right)+\frac{100}{140}\cdot\log\left(\frac{100/150}{100/170}\right)\right) = 1.029454\end{aligned}$$

The cumulative contribution of each new event is shown as the linear line on the log scale in Figure 1.5. The CureVac AG (2020, Table 8) design planned a final analysis at $n = 160$ events, so their implied target was $1.029454^{160} \approx 104$. In comparison to the target score of €104 at 160 events, the actual betting score €1.84 after $83 + 145 = 228$ events in the press release is quite disappointing. Shafer (2021) gives more examples of how betting scores and implied target help to interpret study results in the context of study design.

1.2.3 Honest study design

An implied target does require an honest proposal of the effect of minimal clinical interest $\mu_{\min}$, to evaluate the merits of the study. In regular power analysis, this parameter might be tweaked – e.g. setting an unrealistically large effect – to still argue for the study's advancement with only small sample size. Or the smallest effect size of interest analysis is set after data is observed (Wang et al., 2018). This behavior is incentivized by the

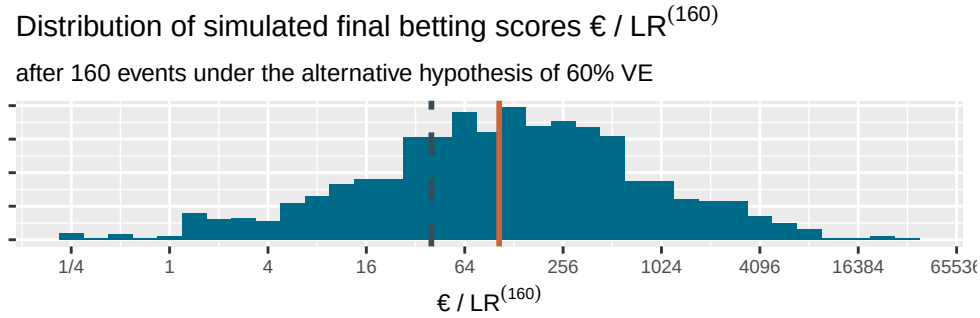

Figure 1.4. *(and Figure 1.5) 1000 simulated sequences of betting scores by round in the FDA game after* 160 *events assuming a probability of* 0.29 *(*40/140*) for each event to occur in the vaccine group. This is the alternative hypothesis of* 60% *VE used to power the CureVac AG (2020) trial at a number of events of 160. The dashed line is the threshold* $1/\alpha = 40$ *one-sided and the solid line is the implied target of €104. Note that the x-axis is on a log scale.*

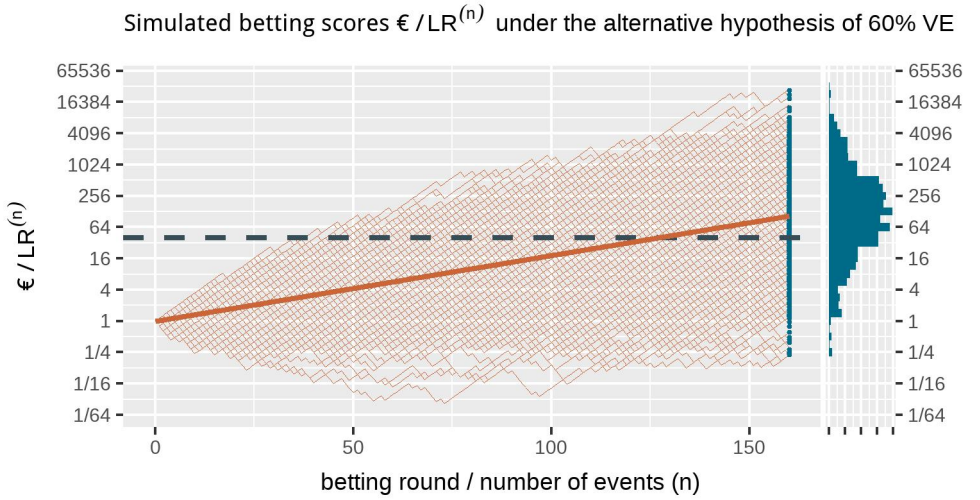

Figure 1.5. *(See above at Figure 1.4.) The histogram for the final betting scores at the right shows the larger scores above and the smaller ones at the bottom, which means that if we turn it, it is the mirror image of the histogram in Figure 1.4. The dashed line is the threshold* $1/\alpha = 40$ *one-sided and the increase in the solid line per additional event/betting round shows the contribution to the implied target at* $n = 160$ *of €104. In this figure, the design has an approximate* 79% *power to observe a betting score/e-value larger than* $1/\alpha = 40$ *before* 160 *events and* 72% *power at exactly* 160 *events (better visible in Figure 1.4). Note that the y-axis is on a log scale.*

all-or-nothing character of Neyman-Pearson tests that also make the power analysis all-or-nothing. If your desired sample size does not meet the power hoped-for, you need to either increase it or abandon the study. This aspect of traditional analyses fully ignores the ideal of cumulative science in which one study is not expected to single-handedly answer a research question and small increments in knowledge are valuable, as long as they build towards a common goal. If they use *e*-values and the ALL-IN framework, researchers do not have to view their analysis as the final one, which helps them to evaluate their study more honestly (Lakens, 2021).

1.3 Collaboration

The *Evidence-Based Research Network* (Lund et al., 2016) aims to always inform new research by past results and to reduce research waste by separating research ideas that are necessary from those that are wasteful. This is not easy to do, however. Different communities might have different notions of necessity or even of what is ethical (a state of so-called *clinical equipoise* (Shamy et al., 2020)). It might therefore be very beneficial to have all those running new clinical trials in a field collaborate together in an ALL-IN meta-analysis.

1.3.1 ALL-IN-META-BCG-CORONA

We ran two ALL-IN meta-analyses during the Covid-19 pandemic with the involvement of seven trials in one and four in the other. All were designed to study whether the BCG vaccine, originally developed to protect against tuberculosis, could protect against Covid-19 (based on a theory of non-specific immune effects and innate immunity (Netea et al., 2020)). The two meta-analyses study different populations (healthcare workers and elderly) and two questions each: the effect of the BCG vaccine on Covid-19 infection (not necessarily symptomatic) and the effect on severe Covid-19 (indicated by hospitalizations). In the following description we will focus on the analysis of Covid-19 infections in the healthcare workers population.

ALL-IN-META-BCG-CORONA followed many of the steps also outlined by Tierney et al. (2021), that we will briefly discuss here: (1) Meta-analysis design, (2) Systematic search for trials, (3) Systematic review for trial inclusion (4) Data upload, and (5) Disseminating results.

(1) Meta-analysis design Early in the project we decided to aim for an IPD meta-analysis on interim data and wrote our protocols and statistical analysis plans. This timestamped two important decisions on the meta-analysis design: the hazard ratio of minimal interest of 0.8 (20% VE) for events of Covid-19 and the level of α set at 0.0025 so the threshold for the *e*-value was at $1/\alpha = 400$. For these decisions we set up a meta-analysis Steering Committee that was still fully blinded to any results at the time. The design was preregistered (Van Werkhoven et al., 2021) and all documentation and a webinar explaining the methodology were made available on a project website (Ter Schure et al., 2020a).

(2) Systematic search for trials We continuously searched for trials to include in the meta-analysis. Some were already known to our Steering committee before we started. They initiated a BCG trial of their own very early in the pandemic and shared their protocol with many of their contacts in the BCG research community. Other trials were found by a repeated systematic search of trial registries. The trials that agreed to join the meta-analysis were each represented by a member in the Advisory Committee. Meetings of the Advisory committee were scheduled regularly and the trials involved could point us to any new developments. A major advantage of ALL-IN meta-analysis here is that the number of trials does not need to be specified in advance.

(3) Systematic review for trial inclusion We received external advice from Cochrane Netherlands on trial inclusion based on a thorough risk-of-bias assessment. For this assessment, each trial shared their protocols, and subsequently all Cochrane's evaluations were shared and discussed with the Steering committee and Advisory committee (where trials were usually represented by their PI's who were blinded to any trial results). Trials had multiple opportunities to answer questions – from Cochrane Netherlands as well as other trials involved – explain their trial and express other concerns about differences between the trials included. The Steering Committee made the final decision on including a trial, before any of that trial's results were known to anyone part of the discussion. The decision of the Steering committee explicitly incorporated both trial quality and meta-analysis homogeneity.

(4) Data upload Parallel to the discussions on trial inclusion, data transfer agreements were signed and data was shared through a secure upload. Each trial had a data uploader that was in close contact with the ALL-IN meta-trial statistician (the author of this thesis) about data quality. The ALL-IN statistician did not attend the discussion meetings and kept the Steering committee and Advisory committee blinded to any results before each trial inclusion decision.

(5) Disseminating results Each data-uploader received a dashboard account with permissions to inspect the meta-analysis e-value and their own trial contribution. Their access of interim meta-analysis results in the dashboard served as a motivator to keep their own trial data upload up-to-date and to check the sequence of e-values for errors. Figure 1.6 shows this dashboard based on a demo login with synthetic data (this demo was available for everyone involved to get an impression). After an initial period where the data-uploaders could only inspect their own trial results, they granted each other permission to inspect all the individual trial contributions. When the first trials were completed and the meta-analysis was approaching its conclusion, the results were also presented to the Advisory and Steering committees. Any interim decisions were planned based on the ALL-IN meta-analysis e-values, but results were also presented in the context of power, implied target per trial and confidence sequences.

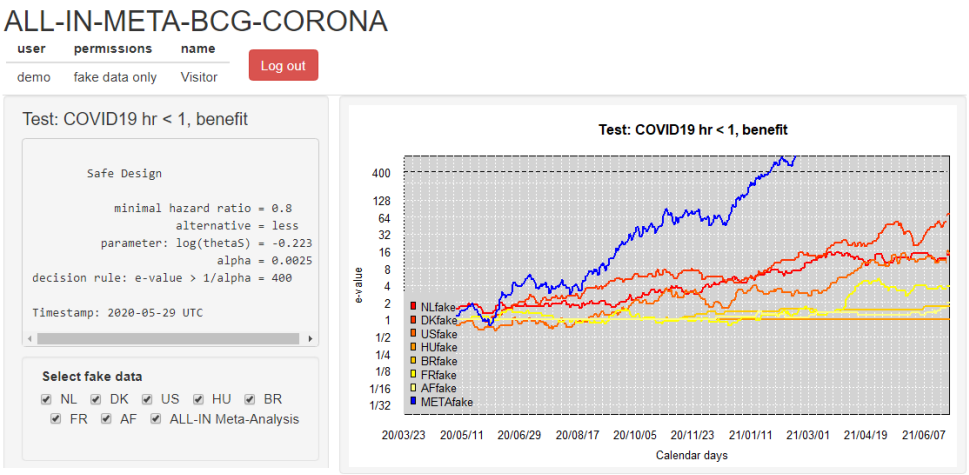

Figure 1.6. *Dashboard used to communicate interim results in* ALL-IN-META-BCG-CORONA *to all data uploaders with a login. The involved trials were performed in the Netherlands (NL), Denmark (DK), the United States (US), Hungary (HU), Brazil (BR), France (FR) and Guinea-Bissau/Mozambique (AF). The dashboard is in demo mode and shows synthetic ("fake") data The option to (de)select trials is for plotting purposes of individual trial e-values; all trials in the dashboard stay included in the meta e-value, following the decision from the Steering committee on trial inclusion. Note that the y-axis is on the log scale.*

1.3.2 Collaborative and bottom-up meta-analysis

In many aspects, our approach agrees with the *Framework for Prospective Adaptive Meta-Analysis (FAME)* from Tierney et al. (2021). FAME argues for prospective meta-analyses in close collaboration with ongoing trials to achieve the same advantages outlined in this chapter, such as aligning trial characteristics ("minimize heterogeneity") and reducing publication bias and "bias [in] both review and meta-analysis methods" introduced by "prior knowledge of trial results" (e.g. accumulation bias (Chapter 3)). In alignment with the FAME recommendations, ALL-IN-META-BCG-CORONA was prospectively designed by preregistering an overall effect size of minimal interest and an α-level. However, ALL-IN meta-analysis in general also allows for a more bottom-up approach when each trial's e-value is based on that trial's own design (effect size of minimal interest, see Section 1.1.4) and trial evidence is synthesized more loosely without a strict decision rule. In comparison to FAME, ALL-IN meta-analysis is much more adaptive. FAME proposes to use conventional meta-analysis (with a fixed sample size) and optimize the timing of the meta-analysis "to anticipate the earliest opportunity for a potentially definitive meta-analysis". In that sense, FAME can only adapt to the speed of recruitment, while ALL-IN allows to adapt to any information so far including the evidence in the trials and the synthesis of the meta-analysis itself. There is also a statistical inconsistency in the FAME approach that is concerned with "striking a balance between maximising the absolute and relative information size and producing a sufficiently timely review" but does explicitly state that

the last step of the meta-analysis is to "assess the value of updating the systematic review and meta-analysis". A fixed sample-size statistical analysis should not be reanalyzed using the same statistical methodology. Any proposal to use conventional meta-analysis for efficiency purposes risks accumulation bias (Chapter 3) because the timing of the meta-analysis might be driven by some of the results part of that same analysis. Hence the best approach is to combine the recommendations from FAME (Tierney et al., 2021) with statistical approach from ALL-IN meta-analysis and the spirit of living systematic reviews.

Collaboration using a dashboard A dashboard for ALL-IN meta-analysis allows us to spot trends in the accumulating evidence, or allow other stakeholders to monitor. A dashboard like Figure 1.6 can give access to the accumulating e-values to those that need to prepare for crossing a threshold in the near future, e.g. for independent data monitoring committees of ongoing trials or for those considering new trials or preparing to update medical guidelines. On a log-scale, the increase in e-values is linear (in expectation) and the observed trends can be estimated, e.g. in Figure 1.6 as an increase in evidence per additional calendar day.

For ALL-IN-META-BCG-CORONA, the time unit t in the definition of $\mathbf{LR}^{\langle t \rangle}$ from (1.4) was set to calendar days and the e-values were updated at each calendar day with an event. The dashboard plots in Figure 1.6 horizontal lines at 1 for trials that do not observe any events yet: they have not started betting and are still at their initial investment of €1 contributing a neutral amount to the multiplication meta-e-value. ALL-IN meta-analysis monitors e-values as events come in, also when they do so from multiple trials simultaneously. In the language of betting, even the analysis of simultaneous events is considered a sequential bet. If the bet on the events from one trial pays out €4, it multiplies our initial capital by 4, and if the events from another trial pay out €5, it does so by a factor 5. Yet if we actually consider those trials to be consecutive bets, we reinvest the €4 from the first into the second, and obtain $€1 \cdot 4 \cdot 5 = €20$, as follows from the definition of the meta-analysis e-value on interim data in (1.4).

Figure 1.6 illustrates what going ALL-IN means: the evidence in all studies can be monitored and compared to the required threshold at any time. The hypothesis test is carried out by comparing the meta e-value in blue to the threshold $1/\alpha$ of 400, plotted as a dotted line. Because the meta-analysis is *anytime* and *live* a conclusion is reached whenever the e-value sequence passes that threshold. The synthesis of studies can efficiently lead the decision to stop recruiting, treat the placebo group or discourage new trials to start, while encouraging inspection of each individual trial's contribution to the meta-analysis. Since each trial's contribution is a simple multiplication, their components can often be conveniently spotted in the agreement of the shape of the meta-analysis and individual trial lines in a dashboard like Figure 1.6 (as long as not too many trials are contributing simultaneously).

Collaboration in a competitive field or a pandemic ALL-IN meta-analysis also prevents losing type-I error control when many trials compete for answers on the same research question, e.g. in an uncoordinated scientific response to a pandemic. If trials are

only evaluated in isolation and a response follows the first positive result of a single trial, serious multiple testing issues arise that inflate the type-I error and result in unreliable inference and, subsequently, poor decisions. This happens especially if all trials perform interim analyses on their own, and a type-I error occurs at an interim analyses before any other trial results are published to refute it. The example dashboard also clearly demonstrates decreased type-II errors: synthesizing the evidence in a meta-analysis at interim stages of the trials, and not after trials are completed, improves the ability to find an effect early. Collaboration is indeed much more efficient.

1.3.3 Fixed-effects and random-effects meta-analysis

Sutton et al. (2007, p. 2491) note that "in a meta-analysis with considerable heterogeneity, the impact of a new (large) study will be (much) less in a random compared to fixed effect model". This is due the incorporation of a parameter in the model that represents the between-study variation. Also Kulinskaya and Wood (2014) find that the goal of sequentially updating a random-effect meta-analysis might involve planning a large number of small trials to estimate the between-study variance well. Even if that is considered advisable, a random-effects model result might still be very difficult to interpret (Riley et al., 2011). Hence there are various reasons to prefer the fixed-effects model to monitor evidence efficiently and to ensure that the trials are sufficiently homogeneous.

Alongside ALL-IN-META-BCG-CORONA we initiated a second ALL-IN meta-analysis. While the first included trials on healthcare workers, the second included trials in the elderly. Early in the process, before seeing any data, our Steering committee noticed that the two groups could be very different. Based on a theory of innate and trained immunity, they expected a different effect of the BCG vaccine on the younger immune system of healthcare workers than on the older immune system in the elderly. It could even be that the BCG vaccine effect was beneficial in the ability to fight off Covid-19 in one population but harmful in the other. In general, the differences between trials can be in three categories: heterogeneous effects, conflicting effect and multiple testing.

Heterogeneous effects Our Steering committee decided that to declare success, all included trials in healthcare workers should observe an effect of 20% VE or larger. If they indeed do, heterogeneity in their effect sizes (e.g. one 20%, one 50%, one 25%) does not matter for their joint ability to reject the *global null hypothesis* of no effect in all trials. So for testing the global null, trials are allowed to be heterogeneous in where they are in the space of the alternative hypothesis $H_1 = \{\text{VE}: \quad 20\% \leq \text{VE} \leq 100\%\}$. For estimation, however, it is not clear what the ALL-IN confidence interval is estimating if we assume that the effects in the trials are very different. Still, as a first summary, a *typical effect size* (Peto, 1987) might be useful if we are unable to estimate a random effects model. The development of confidence sequences for random-effects meta-analysis is a major goal for future work. We do not, however, believe that the evidence in a line of research should be monitored based on whether this interval excludes the null hypothesis, or whether the e-value corresponding to the random-effects null model does: for testing, the global null is much more natural. Waiting for a random-effect model to reach a certain threshold

is counter-intuitive, since it might require many small trials to estimate the between-trial variability instead of focusing on testing the treatment effect. Moreover, the goal of rejecting the null hypothesis corresponding to this model can be quite strange. When testing a zero-effect null hypothesis, it assumes that there are true effects of harm and true effects of benefit among the trials and that their mean is exactly zero.

Conflicting effects If one of the trials has an effect smaller than 20% or even a harmful effect, we should anticipate betting scores or e-values that are smaller than 1. So a meta-analysis multiplication of those e-values would reduce the evidence available from other trials. If we can identify groups for which we expect that the trials in each group have an effect in the same direction and of at least the minimal size, we can perform separate meta-analyses. This was the rationale behind grouping healthcare workers and the elderly each in their own ALL-IN-META-BCG-CORONA analysis.

Multiple testing When our analysis is exploratory, and we really have no idea how to group the various trials, we are faced with a multiple testing problem. Note that in this situation also no conventional meta-analysis method would be used to test a common null-hypothesis. We wonder whether any of the trials has the ability to reject the null hypothesis. In that case, we can divide our initial investment over the trials, and see if the totality of their bet achieves a high betting score. Research into this use of e-values has shown that indeed averaging e-values is the optimal way to have type-I error control in a standard multiple testing setting (Vovk and Wang, 2021). We return to the notion of hedging bets and averaging e-values in Section 1.4.

Problems with heterogeneity in meta-analysis are not tied to the ALL-IN approach and familiar to anyone working with meta-analysis methods. ALL-IN-META-BCG-CORONA had the advantage that many of the trials that started later had drawn inspiration from the protocol of the first trial. The same sort of alignment of inclusion criteria and outcome definitions might be achieved in other lines of research as well. Hence close collaboration can be very important and the promise of an early conclusion of the research effort might keep a research field motivated to keep the goals aligned.

1.4 Communication

We have illustrated that the language of betting can be useful in interpreting results from an ALL-IN meta-analysis. Here we argue this further by giving extensions of our method that are very easily explained in terms of betting.

1.4.1 The language of betting for two-sided tests

Our examples so far covered one-sided tests, but those can be easily extended to two-sided tests, e.g. by taking

$$\mathbf{LR}^{(n)}_{\text{two-sided}} = \frac{1}{2} \cdot \left(\mathbf{LR}^{(n)}_{\text{left}} + \mathbf{LR}^{(n)}_{\text{right}} \right),$$

with

$$\mathbf{LR}^{(n)}_{\text{left}} = \frac{\phi_{\mu_{\min(\text{left})}\sqrt{n}}(z^{(n)})}{\phi_{\mu_0}(z^{(n)})} \quad \text{and} \quad \mathbf{LR}^{(n)}_{\text{right}} = \frac{\phi_{\mu_{\min(\text{right})}\sqrt{n}}(z^{(n)})}{\phi_{\mu_0}(z^{(n)})},$$

to represent a two-sided alternative hypothesis

$$H_1 = \left\{ \phi_{\mu_1} : \quad \mu_1 \leq \mu_{\min(\text{left})} \quad \text{or} \quad \mu_1 \geq \mu_{\min(\text{right})} \right\}.$$

Such a two-sided test is easy to interpret in the language of betting. We essentially split our initial investment (e.g. €1) between the two sides of the alternative hypothesis (e.g. by betting €0.50 on one side and €0.50 on the other). Any other weighting of the two sides is also possible and corresponds to a different division of the initial investment. The crucial thing is that each side tests the same null hypothesis $H_0 = \{\phi_{\mu_0}\}$ and has expectation 1 under the null hypothesis, such that any weighted average also has expectation 1 and is an e-value. Note that for a meta-analysis at time t with $k\langle t \rangle$ studies this becomes:

$$\mathbf{LR}^{\langle t \rangle}_{\text{two-sided}} := \frac{1}{2} \left(\prod_{i=1}^{k\langle t \rangle} \mathbf{LR}^{(n_i\langle t \rangle)}_{i,\text{left}} + \prod_{i=1}^{k\langle t \rangle} \mathbf{LR}^{(n_i\langle t \rangle)}_{i,\text{right}} \right). \tag{1.9}$$

Usually one side of the bet is losing and the other is winning such that we do not want to reinvest (multiply) across sides but keep them separate for all trials. In our ALL-IN-META-BCG-CORONA dashboard we also visualized these two sides of the meta-analysis test separately; in Figure 1.6 we show only the left-sided test (for benefit) of the two.

1.4.2 The language of betting for co-primary endpoints

Another way to hedge our bets is by considering multiple primary outcomes. In ALL-IN-META-BCG-CORONA, for example, not only the Covid-19 events were counted, but Covid-19 hospitalizations as well, as an indicator for severe disease. We started with $\alpha = 0.05$ and put 10% on Covid-19 ($\alpha = 0.0025$ on each of the two sides of a two-sided test) and 90% on hospitalisations ($\alpha = 0.0225$ on each of the two sides of a two-sided test). So the thresholds to achieve with the e-value for Covid-19 was set at $1/\alpha = 400$ and the one for hospitalization at $1/\alpha = 44.44$. A different way to formulate this is that each had to achieve $1/\alpha = 20$, but that the sequence of e-values for Covid-19 started with an initial investment of €0.05 for each side of the two-sided test (and had to multiply by 400 to reach €20) and that the e-value for hospitalization started with an initial investment of €0.45 for each side (and had to multiply by 44.44 to reach €20).

There are two ways to consider such a bet on two co-primary outcomes: separately and combined. If we evaluate the e-values for each primary outcome separately and reach the

threshold with either of the two, we are rejecting the null for that outcome. We are doing two separate tests. If we evaluate the e-values combined, we average them weighted by their α, just as for the two sides of the two-sided test. In that case we have similar type-I error control, but reject the null hypothesis that both are a null effects in favor of the alternative hypothesis that one of them is not. Yet we cannot conclude which one is non-null with the same type-I error since our α level applies to the combined bet and the individual components to the averaged bet are essentially lost.

1.5 Concluding remarks

The novelty of this chapter lies in a new method for meta-analysis. We do not claim any novelty for the underlying mathematics, though. The basic methods we describe can be viewed as relatively minor variations of the anytime-valid tests that are designed to preserve type-I error under optional stopping, as designed by H. Robbins and his students (Darling and Robbins, 1968; Robbins, 1970). Unfortunately and surprisingly, these tests have not caught on in statistics until a few years ago – right now they are thriving in work on so-called *safe tests*, *anytime-valid confidence sequences* and e-values e.g. Shafer et al. (2011); Johari et al. (2021); Pace and Salvan (2019); Howard and Ramdas (2019); Howard et al. (2021); Ramdas et al. (2020); Vovk and Wang (2021); Shafer (2021); Grünwald et al. (2019); Turner et al. (2021); Henzi and Ziegel (2021). As far as we know, it has never before been suggested to use such methods in a meta-analysis context. (Group sequential methods, which have originally also been inspired by the anytime-valid tests, have in turn spurred developments in meta-analysis, but these are substantially different from ALL-IN.) Also, the fact that the logrank test can give a likelihood ratio of the type needed for an anytime-valid test/an ALL-IN meta-analysis is a new finding described in Chapter 2.

Likelihood ratios, E-variables and e-values

In this chapter we presented betting scores/e-values that are equivalent to likelihood ratios. In general though, betting scores and e-values are really generalizations of likelihood ratios that preserve the properties of likelihood ratios that give them a prominent role in statistics. Entire books have been written to advocate for summarizing evidence in observed data by a likelihood ratio (Edwards, 1974; Royall, 1997) and to separate the goal of measuring evidence from expressing posterior beliefs and making decisions. Likelihood ratios have the property that they can "favor a true hypothesis over a false one more and more strongly" and while a likelihood ratio can be misleading, "strong evidence cannot be misleading very often" (Royall, 1997, p. 14). This latter type-I error control is also referred to as a *universal bound* by Royall (1997) and, by recognizing Ville's inequality, can be generalized to other betting scores and e-values.

A betting score $\in$ is a random outcome of a bet and its random variable is an E-variable if it is nonnegative and for all $P \in H_0$, $\mathbf{E}_P[\in] \leq 1$. For a given outcome of the bet, the value of such a random variable is the betting score or e-value. Ville's inequality relies on the multiplication of E-variables – forming a test martingale – which also has expectation smaller than 1 and thus is itself an E-variable. For the example e-values in this chapter,

the requirement on the expectation $\mathbf{E}_0[\mathbf{LR}] \leq 1$ holds for a simple null hypothesis, e.g. $H_0 = \{\phi_0\}$.

Apart from likelihood ratios of two simple hypothesis, e-values can also be defined for more complicated tests – e.g. a t-test with a nuisance parameter for the variance – in which case the unit expectation needs to hold not for a single mean-0-normal distribution with known variance, but for all mean-0-distributions with any variance. Grünwald et al. (2019) shows that it often is possible to construct E-variables for such composite testing problems, which is why we consider the e-value the right generalization of the likelihood ratio.

Anytime-valid confidence sequences

In this chapter we briefly presented a confidence sequence (in Figure 1.3) for the hazard ratio or VE that was based on the Gaussian approximation to the logrank statistic and the Peto (1987) estimator. This estimator can be derived from summary statistics and is therefore still quite common in meta-analysis as a so-called *two-stage method*, although it is advised against for extreme hazard ratios (Simmonds et al., 2011). Research into other confidence sequences for the hazard ratio is still ongoing. For other estimation problems, confidence sequences already have been thoroughly studied, for example for medians and other quantiles (Howard and Ramdas, 2019), and odds ratios (Turner et al., 2021). These have not, however, been extended to meta-analysis, and especially for the random-effects meta-analysis model, research into confidence sequences is a major goal of future work.

Availability in software

The safestats R package (Turner et al., 2022) provides software to do an e-value analysis for the t-test, Z-test, logrank test and $2x2$-tables. Also functions are available to calculate the power and implied target for these study designs. Confidence sequences can be calculated for the odds ratio in 2x2-tables and the hazard ratio in time-to-event data.

Competing interests

Both authors are proud to have received three shots of the Pfizer-BioNTech Covid-19 vaccine. No further competing interests were disclosed.

Appendices

1.A The inverse-conservative *p*-value

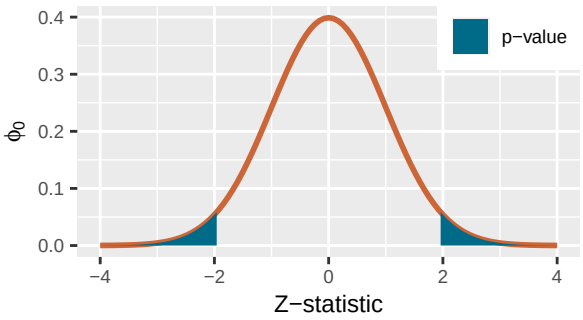

Figure 1.A.7. Two-sided Z-test against the null hypothesis $H_0 = \{\phi_0\}$ observing $z = 1.96$ with a p-value of 0.05.

The standard way to introduce p-values in introductory texts is to use a graph like Figure 1.A.7 that shades the tail area of a sampling distribution under the null hypothesis for an observed test statistic. In Figure 1.A.7 the observed test statistic is a Z-statistic of 1.96, such that the two-sided p-value is 0.05. The precise definition of this p-value for the random variable Z and any observation z (e.g. $z = 1.96$) is:

$$p\text{-value}(z) = \mathbf{P}_{Z\sim\phi_0}[|Z| \geq z] \tag{1.A.1}$$

$$\text{such that for all } 0 \leq \alpha \leq 1\text{:} \quad \mathbf{P}_0[p\text{-value} \leq \alpha] = \alpha, \tag{1.A.2}$$

with $\mathbf{P}_0$ the probability under the null hypothesis. While explaining-p-by-picture is helpful at first, it also has strong limitations. For example, if the sample size n is not fixed in advance but determined by a fixed, a priori known stopping rule (e.g. stop as soon as the first z comes in that has value > 2.5), then n will itself be random and the threshold for the test statistic will depend on n; there is no evident visualization.[4] For this reason, it is common in theoretical statistics to define p-values directly as random variables that have a uniform distribution under the null. From (1.A.1) and (1.A.2) we see that this is compatible with the introductory approach.

Conservative *p*-values Defined by property (1.A.2), the standard (strict) p-value can be generalized to a (*conservative*) p-value that does not have a uniform distribution, but

[4] An exception is the case in which there are just two possible sample sizes, n_1 and n_2, with the rule for choosing between n_1 and n_2 based on the first n_1 data points $z^{(n_1)}$ given (e.g. stop if $z^{(n_1)} > 2.5$, continue otherwise). We can then specify a bivariate distribution and evaluate the tail area for the combined observations (e.g. $(z^{(n_1)}, z^{(n_2)})$). This is sometimes illustrated in introductory texts on group-sequential methods, but cannot be extended beyond this (3D) bivariate case.

whose distribution is stochastically dominated by the uniform distribution under the null hypothesis. This means that the equality in equation (1.A.2) is replaced by an inequality:

$$\mathbf{P}_0[p\text{-value} \leq \alpha] \leq \alpha. \tag{1.A.3}$$

For such conservative p-values, it still holds that, for any fixed significance level α, the test that rejects if $p \leq \alpha$ has type-I error at most α. The inequality is thus in the right direction to preserve type-I error guarantees.

Our first observation is very simple: for any fixed n, the likelihood ratio $\mathbf{LR}^{(n)}$ is an inverse-conservative p-value for samples of that size n, i.e. (1.A.3) holds with 'p-value' set to $1/\mathbf{LR}^{(n)}$. However, simply viewing **LR**s as inverse conservative p-values is not nearly doing them sufficient justice. A better (yet still incomplete) comparison is to 'anytime-valid p-values', a terminology that stems from Johari et al. (2021). Whereas standard p-values only make sense for an a priori given, fixed sample size n or a priori given, fixed stopping rule, anytime-valid p-values are really sequences $p'_1, p'_2, \ldots$, one for each sample size n. Even if the stopping rule for determining n is unknown in advance, or unclear, or greedy ('stop at first n such that $p_n \leq \alpha$'), the type-I error guarantee is preserved. It turns out that the inverse of the **LR**s as defined for ALL-IN meta-analysis can be viewed in this way: the sequence $(1/\mathbf{LR}^{(1)}, 1/\mathbf{LR}^{(2)}, \ldots)$ as defined in Section 1.1 is a sequence of anytime-valid p-values.

1.B R Code for calculations, simulations and plots

R code for the calculations, simulations and plots in this chapter can be found on the Open Science Framework (Ter Schure, 2021b, `https://osf.io/d9jny/`), including:

- Settings of the FDA game
- Pfizer/BioNtech results in the FDA game
- CureVac results in terms of confidence interval and the FDA game
- Plotting the FDA game betting scores under the null hypothesis (Figure 1)
- Plotting the expected sample size for various strategies in the FDA game (Figure 2)
- Plotting the anytime-valid confidence sequence for a random ordering of the CureVac data (Figure 3)
- Plotting the implied target of the CureVac design in the FDA game (Figure 4 & 5)
- Plotting a p-value of 0.05 (Figure 6; Appendix)

2 | The Safe logrank test

Abstract

We introduce the safe logrank test, a version of the logrank test that provides type-I error guarantees under optional stopping and optional continuation. The test is sequential without the need to specify a maximum sample size or stopping rule and allows for cumulative meta-analysis with type-I error control. The method can be extended to define anytime-valid confidence intervals. All these properties are a virtue of the recently developed martingale tests based on E-variables, of which the safe logrank test is an instance. We demonstrate the validity of the underlying nonnegative martingale in a semi-parametric setting of proportional hazards and show how to extend it to ties, Cox' regression and confidence sequences. Using a Gaussian approximation on the logrank statistic, we show that the safe logrank test (which itself is always exact) has a similar rejection region to O'Brien-Fleming α-spending but with the potential to achieve 100% power by optional continuation. Although our approach to *study design* requires a larger sample size, the *expected* sample size is competitive by optional stopping.

Introduction

Traditional hypothesis tests and confidence intervals lose their type-I error and coverage guarantees, thus, validity and interpretability, under *optional stopping* and *continuation*. Roughly, optional stopping refers to stopping earlier than originally planned, for example, when results look good enough; optional continuation refers to adding additional data at the end of a trial, or even starting a new trial and combining results, for example, when results of the trial are promising but not fully conclusive.

Recently, a new theory of testing and estimation has emerged for which optional stopping and continuation pose no problem at all (Shafer et al., 2011; Howard et al., 2021; Ramdas et al., 2020; Vovk and Wang, 2021; Shafer, 2021; Grünwald et al., 2019; Turner et al., 2021). The main ingredients are (1) the E-variable that has as outcome an e-value, a direct alternative to the classical p-value, and (2) the test martingale, a product of conditional E-variables. Both are used to create so-called *safe* tests that provide type-I error control under optional stopping and optional continuation, and *anytime-valid* confidence intervals that remain valid irrespective of the stopping time employed. Pace and Salvan

(2019) argue that even without optional stopping, anytime-valid confidence intervals may be preferable over standard ones. Here we provide a concrete instance of this theory: we develop E-variables and martingales for a safe version of the classical logrank test of survival analysis (Mantel, 1966; Peto and Peto, 1972): safe under both optional stopping and continuation. The E-variables, martingales and the corresponding tests are implemented in the SafeStats R package (Turner et al., 2022).

The logrank test is often used in randomized clinical trials to test a difference between two groups in survival time or other time to an event. Its test statistic appears as the score test corresponding to the Cox (1972) proportional hazards model – when the only covariate is the treatment/control indicator – and is also a key tool in statistical monitoring of trials by means of group sequential/α-spending approaches. These sequential methods allow several interim looks at the data to stop for efficacy or futility. Like ours, they are connected to early work by H. Robbins and his students (Darling and Robbins, 1967; Lai, 1976), but the details are very different and in some cases, our approach is more straightforward. In case of unbalanced allocation, for example, α-spending approaches do not provide strong type-I error guarantees (Wu and Xiong, 2017) due to the approximations involved. The basic version of the safe logrank test, however, is exact, without any approximations, so that unbalanced allocation is no problem at all. This ties to the important advantage of using E-variables instead of α-spending: it is more flexible, and as a consequence, easier to use.

Group sequential approaches require prespecified interim looks, in terms of the information fraction of the trial; α-spending is less rigid but still needs a maximum sample size to be set in advance. These requirements limit the utility of a promising but non-significant trial once the maximum sample size is reached, because extending such a trial makes it impossible to control the type-I error. Moreover, also new trials cannot be added in a typical retrospective meta-analysis (not prespecified before any trial), when the number of trials or timing of the meta-analysis are dependent on the trial results. Such dependencies introduce accumulation bias and invalidate the assumptions of conventional statistical procedures in meta-analysis (Chapter 3). In contrast, an analysis based on E-variables can extend existing trials as well as inform whether to do new trials and meta-analyses, while still controlling type-I error rate. Type-I error control is retained even (i) if the e-value is monitored continuously and the trial is stopped early whenever the evidence is convincing, (ii) if the evidence of a promising trial is increased by extending the experiment and (iii) if a trial result spurs a new trial with the intention to combine them in a meta-analysis. Even with dependence between the trials, the test based on the multiplication of these e-values retains the type-I error control, as long as all trials test the same (i.e., global) null hypothesis. This becomes especially interesting if we want to combine the results of several trials in a bottom-up retrospective meta-analysis, where no top-down stopping rule can be enforced. We can even combine interim results of trials by multiplication while these trials are still ongoing – going beyond the realm of traditional sequential approaches.

Contributions and content We show that Cox' partial likelihood underlying his proportional hazards model can be used to define E-variables and test martingales. We first do this for (a) the case with only a group indicator (no other covariates) and without simultaneous events (ties) in Section 2.1.1, leading to a logrank test that is safe for optional stopping. We extend this to (b) the case with ties in Section 2.1.2 and to (c) a Gaussian approximation on the logrank statistic in Section 2.1.3 that is useful if only summmary statistics are available. We provide extensive computer simulations in Section 2.2 and Section 2.3, comparing our 'safe' logrank test to the traditional logrank test and α-spending approaches. In Section 2.2 we show that the exact safe logrank test has a similar rejection region to O'Brien-Fleming α-spending for those designs and hazard ratios where it is well-approximated by a Gaussian safe logrank test (case (c)). While always needing a bit more data in the design phase (the price for indefinite optional continuation), the expected sample size needed for rejection remains very competitive. We might want to design for a maximum sample size to achieve a certain power, but need a smaller sample size on average since we can safely engage in optional stopping. In Section 2.4 we extend the approach in various directions: first, in its basic version, the safe logrank test requires specification of a minimum clinically relevant effect size. If instead one wants to learn the actual effect size of the data and/or infuse prior knowledge about the effect size into the method via a Bayesian prior, this can be done without any difficulties. The resulting version of the safe test keeps providing non-asymptotic frequentist type-I error control even if these priors are wildly misspecified (i.e., they predict very different data from the data we actually observe); this is discussed in Section 2.4.1. We then show how our logrank test can be inverted to allow for *anytime-valid* confidence sequences (Section 2.4.2), and we provide the extension to covariates (Section 2.4.3). This extension, based on the Cox model, requires solving a complicated optimization problem and implementation is therefore deferred to future work.

To keep the exposition simple, when introducing our methods we represent data by a simplified discrete-time stochastic process, in which our test statistics take the form of likelihood ratios. In Appendix Section 2.A we show how our test statistics remain valid E-variables and test martingales under a proportional hazard assumption in continuous time; our likelihood ratios then become partial likelihood ratios. Once the definitions are in place, these results are mostly straightforward consequences from earlier work, in particular (Cox, 1972; Slud, 1992; Andersen et al., 1993). The novelty of our work is thus mainly in *defining* the new tests in the first place and showing by computer simulation that, while being substantially more flexible, they show competitive behavior with existing approaches, i.e. the classical logrank test in the fixed design setting and in combination with α-spending.

We delegate to the Appendix insights that, while important, are not needed to follow the main development. Most importantly, the particular E-variable we design satisfies the GROW criterion. Grünwald et al. (2019) provide several motivations for this criterion. We provide an additional one in Appendix Section 2.B by an argument (originally due to Breiman (1961), but not widely known) showing that it leads to tests with minimal expected stopping time. In the remainder of this introduction, we provide a short introduction to E-variables, test martingales and safe tests.

E-Variables, Test Martingales, Safety and Optimality

In this subsection we briefly introduce general concepts necessary to develop E-variables for the logrank test. The concepts are borrowed from Grünwald et al. (2019) (GHK from now on), which provides an extensive introduction to E-variables and its relation to likelihood ratios and Bayes factors. As seen in Example 1 below, when both the null H_0 and alternative hypotheses H_1 are simple, the most familiar E-variable is the likelihood ratio itself, where the product of likelihood ratios is also an E-variable. In fact, it is the best E-variable in the GROW sense defined further below. Similarly, when H_0 and or H_1 are composite, E-variables are often, but certainly not always, Bayes factors; and Bayes factors are certainly often, but not always E-variables. E-variables also have a crisp interpretation in terms of betting scores (GHK, Shafer (2021)). Briefly, a test martingale, which is the running product of a sequence of E-variables, describes the total profit you have made so far in a sequential gambling game in which you would not expect to win any money if the null were true. As a consequence, as expressed by (2.3) below, when the null holds true, there is little chance for the martingale to *ever* take on a large value. The general story that emerges from papers such as Shafer's as well as GHK and Ramdas et al. (2020) is that E-variables and test martingales are the 'right' generalization of likelihood ratios to the case that either or both H_0 and H_1 are composite. Existing tests based on generalizations of likelihood ratios to composite testing problems that are not E-variables often show problematic behavior in terms of nonasymptotic error control, whereas those generalizations that are E-variables can be combined freely over experiments while retaining type-I error control, thereby providing an intuitive notion of evidence.

Definition 2.0.1. *Let S be a nonnegative random variable defined on a sample space Ω, and let H_0, the null hypothesis, be a set of distributions for Ω. We call S an E-variable if for all $P \in H_0$, $\mathbf{E}_P[S] \leq 1$. For arbitrary random variable Z on Ω, S is called an E-variable* conditional on Z *if for all $P \in H_0$, $\mathbf{E}_P[S \mid Z] \leq 1$. A (conditional) E-variable is called* sharp *if for all $P \in H_0$, the inequality holds as an equality.*

We now consider not a fixed sample space, but a sequence of samples.

Definition 2.0.2. *Let $Y\langle 1\rangle, Y\langle 2\rangle, \ldots$ represent a discrete-time random process and let H_0, the null hypothesis, be a collection of distributions for this process. Fix $i > 0$ and let $S\langle i\rangle$ be a nonnegative random variable that is determined by (i.e. can be written as a function of) $(Y\langle 1\rangle, \ldots, Y\langle i\rangle)$. As an instance of Definition 2.0.1, for*[1] *$i \geq 0$, we say that $S\langle i+1\rangle$ is an E-variable conditionally on $(Y\langle 1\rangle, \ldots, Y\langle i\rangle)$ if for all $P \in H_0$,*

$$\mathbf{E}_P\left[S\langle i+1\rangle \mid Y\langle 1\rangle, \ldots, Y\langle i\rangle\right] \leq 1. \tag{2.1}$$

Definition 2.0.3. *If, for each i, $S\langle i\rangle$ is an E-variable conditional on $Y\langle 0\rangle, \ldots, Y\langle i-1\rangle$, then we say that the product process $S^{\langle 1\rangle}, S^{\langle 2\rangle}, \ldots$ with $S^{\langle n\rangle} := \prod_{i=1}^{n} S\langle i\rangle$ is a* test supermartingale *relative to the given H_0. If all constituent E-variables are sharp, we call the process a test martingale.*

[1] For the case $i = 0$, (2.1) should be read as $\mathbf{E}_P\left[S\langle 1\rangle\right] \leq 1$.

We note that, for a supermartingale $S^{\langle 1\rangle}, S^{\langle 2\rangle}, \ldots$, each $S^{\langle i\rangle}$ is itself an (unconditional) E-variable.[2]

Example 1. [(Partial) Likelihood Ratios as E-Values and Test Martingales] Let $H_0 = \{P_0\}$ and $H_1 = \{P_1\}$ both be simple, each containing a single distribution for the process $Y\langle 1\rangle, Y\langle 2\rangle, \ldots$ with each $Y\langle i\rangle$ taking values in a finite set $\mathscr{Y}$. Let, for each $i \in \mathbb{N}$, $p_0(Y\langle i+1\rangle \mid Y\langle 1\rangle, \ldots, Y\langle i\rangle)$ and $p_1(Y\langle i+1\rangle \mid Y\langle 1\rangle, \ldots, Y\langle i\rangle)$ be the conditional probability mass functions corresponding to P_0 and P_1. Then $S\langle i+1\rangle := p_1(Y\langle i+1\rangle \mid Y\langle 1\rangle, \ldots, Y\langle i\rangle)/p_0(Y\langle i+1\rangle \mid Y\langle 1\rangle, \ldots, Y\langle i\rangle)$ is a sharp E-variable, as is seen from calculating (2.1) (an explicit calculation is given for the logrank test in (2.9)). Further, $S^{\langle n\rangle} = \prod_{i=1}^{n} p_1(Y\langle i\rangle \mid Y\langle 1\rangle, \ldots, Y\langle i-1\rangle)/p_0(Y\langle 1\rangle, \ldots, Y\langle i\rangle \mid Y\langle i-1\rangle) = p_1(Y\langle 1\rangle, \ldots, Y\langle n\rangle)/p_0(Y\langle 1\rangle, \ldots, Y\langle n\rangle)$ is the likelihood ratio for the first n outcomes, and the test supermartingale $S^{\langle 1\rangle}, S^{\langle 2\rangle}, \ldots$ is a likelihood ratio process.

More generally, let $Y\langle 1\rangle, Y\langle 2\rangle, \ldots$ be a discrete stochastic process defined on an arbitrary underlying measurable space (which may e.g. represent time as a continuous-valued random variable) and let H_0 be any set of probability distributions on this space such that all elements in H_0 agree on the conditional probability mass functions $p_0(Y\langle i+1\rangle \mid Y\langle 1\rangle, \ldots, Y\langle i\rangle))$ for $i = 1, 2 \ldots$. Let $q(Y\langle i+1\rangle \mid Y\langle 1\rangle, \ldots, Y\langle i\rangle)), i = 1, 2, \ldots$ and $r(Y\langle i+1\rangle \mid Y\langle 1\rangle, \ldots, Y\langle i\rangle)), i = 1, 2, \ldots$ denote any other sequence of conditional probability mass functions for this discrete process. Then we have, by the same explicit calculation, that

$$S\langle i+1\rangle := \frac{r(Y\langle i+1\rangle \mid Y\langle 1\rangle, \ldots, Y\langle i\rangle)}{q(Y\langle i+1\rangle \mid Y\langle 1\rangle, \ldots, Y\langle i\rangle)} \text{ is a sharp conditional } E\text{-variable}$$
$$\text{if } p_0(Y\langle i+1\rangle \mid Y\langle 1\rangle, \ldots, Y\langle i\rangle)) = q(Y\langle i+1\rangle \mid Y\langle 1\rangle, \ldots, Y\langle i\rangle)). \tag{2.2}$$

If this is indeed the case for each i, then $S^{\langle 1\rangle}, S^{\langle 2\rangle}, \ldots$ with $S^{\langle n\rangle} = \prod_{i=1}^{n} r(Y\langle i\rangle \mid Y\langle 1\rangle, \ldots, Y\langle i-1\rangle)/p_0(Y\langle i\rangle \mid Y\langle 1\rangle, \ldots, Y\langle i-1\rangle) =$
$r(Y\langle 1\rangle, \ldots, Y\langle n\rangle)/p_0(Y\langle 1\rangle, \ldots, Y\langle n\rangle)$ is a test martingale and $S^{\langle n\rangle}$ is now in general not a full, but so-called *partial* likelihood — but the difference will not matter for our purposes.

For the special case of likelihood ratios, the following safety result is referred to as a *universal bound* by Royall (1997).

Safety The interest in E-variables and test martingales derives from the fact that we have type-I error control irrespective of the stopping rule used: for any test (super-) martingale $\{S^{\langle i\rangle}\}_{i\in\mathbb{N}}$ relative to $\{Y\langle i\rangle\}_{i\in\mathbb{N}}$ and H_0, Ville's inequality (Ville, 1939; Shafer et al., 2011) tells us that for all $0 < \alpha \leq 1$, $P \in H_0$,

$$P(\text{there exists } i \text{ such that } S^{\langle i\rangle} \geq 1/\alpha) \leq \alpha. \tag{2.3}$$

In other words, under the null there is little chance (e.g., less than or equal to 5%) that a supermartingale $S^{\langle i\rangle}$ will *ever* realize a large value (e.g., larger than 20). Ville's inequality

[2] Readers familiar with measure-theoretic terminology may recognize a test (super-) martingale as a non-negative (super-) martingale relative to the filtration $\sigma(Y\langle 1\rangle), \sigma(Y\langle 1\rangle, Y\langle 2\rangle), \ldots$. The requirement that '$S\langle i\rangle$ is determined by $(Y\langle 1\rangle, \ldots, Y\langle i\rangle)$' then means that the process $\{S\langle i\rangle\}_{i\in\mathbb{N}}$ is adapted to this filtration.

implies that type-I error control is guaranteed regardless of the stopping rule, which may be deterministic, e.g., stop after 100 samples, or data-driven, e.g., stop as soon as $S^{\langle i \rangle} \geq 20$, or may be exogenuously determined, e.g. stop when your boss tells you to.

Thus, if we measure evidence against the null hypothesis after observing i data units by $S^{\langle i \rangle}$, and we reject the null hypothesis if $S^{\langle i \rangle} \geq 1/\alpha$, then our type-I error will be bounded by α, no matter what stopping rule was used for determining i. We thus have type-I error control if we run out of data, or money to gather new data, and even if we use the most aggressive stopping rule compatible with this scenario: stop at the first i at which $S^{\langle i \rangle} \geq 1/\alpha$. We also have type-I error control if the actual stopping rule is unknown to us, or determined by external factors independent of the data $Y\langle i \rangle$ – as long as the decision whether to stop depends only on past data, and not on the future. This is impossible to achieve with standard Neyman-Pearson tests.

We will call any procedure that takes as input stopping time τ, significance level α and the sequence of random variables $S\langle 1 \rangle, S\langle 2 \rangle, \ldots$, that stops at (random) time τ and that outputs REJECT if $S^{\langle \tau \rangle} \geq 1/\alpha$ and $accept$ otherwise, a *level α-test that is safe under optional stopping*, or simply a *safe test*.

Importantly, we can also deal with *optional continuation*: we can combine E-variables from different trials that share a common null (but may be defined relative to a different alternative) by multiplication, and still retain type-I error control – see Example 3. If we used p-values rather than E-variables we would have to resort to e.g. Fisher's method, which, in contrast to multiplication of e-values, is invalid if there is a dependency between the (decision to perform) tests. Finally, E-variables and test martingales can also be used to define 'anytime-valid confidence intervals' that remain valid under optional stopping (Section 2.4.2).

Optimality Just like p-values, E-variables only require the specification of a null model H_0. Provided with such a null model, we can typically specify a large class of E-variables, denoted here by $\mathscr{S}$, all which remain small with high probability under the null. If, on the other hand, the data are governed by a distribution belonging to a given alternative model H_1, then we would like to choose that E-variable from $\mathscr{S}$ that accumulates evidence against the null as fast as possible. The speed of evidence accumulation under the alternative is defined (conservatively) as the smallest expectation of the logarithm of the E-variable under the alternative, see GHK and Shafer (2021) for various reasons for this choice. More specifically, provided with an alternative H_1, the growth rate of a conditional E-variable $S\langle i \rangle$ under distribution $P_1 \in H_1$ is defined as $\mathbf{E}_{P_1}[\log S\langle i+1 \rangle \mid Y\langle 1 \rangle, \ldots, Y\langle i \rangle]$. The optimal E-variable amongst all E-variables conditional on $Y\langle 1 \rangle, \ldots, Y\langle i \rangle$ is that S that solves the criterion

$$\max_{S} \min_{P \in H_1} \mathbf{E}_P[\log S\langle i+1 \rangle \mid Y\langle 1 \rangle, \ldots, Y\langle i \rangle]. \tag{2.4}$$

We call the optimal E-variable GROW, i.e., growth-rate optimal in worst-case. Appendix Section 2.B provides another motivation for using the logarithm to measure the speed of evidence accumulation, originally provided by Breiman (1961). Specifically, we show

that, under the alternative, the GROW E-variable minimizes the expected number of data points needed to reject the null if no maximum sample size is specified. Note that this is analogous to finding a test that maximizes power. In Section 2.3 we provide some simulations to relate power to GROW. Note that we cannot directly use power in designing tests, since the notion of power requires a fixed sampling plan, which by design we do not have.

2.1 Safe logrank tests

The classical logrank test compares the risk of an event in two groups (e.g. a treatment and a control group) in a nonparametric test, meaning that it requires no underlying (parameterized) distribution on the event times. No matter when the events occur, the null hypothesis assumes that their probability is equal for all participants, regardless of their group. We can flip this null hypothesis on its head for each event time: given that an event has occurred, it has equal probability to have been in either the treatment or the control group, when an equal number of participants are at risk in both groups; more generally, the probability ratio between an event in both groups is equal to the ratio of participants in these groups. That risk set (of participants at risk) changes with each event so the test crucially relies on the order (the ranks) of the events.

Logrank hypotheses tested We observe a sequence of event times $t\langle 1\rangle < t\langle 2\rangle < t\langle 3\rangle < \dots$ such that for all i, at time $t\langle i\rangle$, one event happens, and inbetween $t\langle i\rangle$ and $t\langle i+1\rangle$, no events happen. The time until the i^{th} event occurs – $t\langle i\rangle$ itself – does not play a role in the logrank statistic; only the ordering of events matters. The hypothesis tested by the logrank test can be expressed in terms of the instantaneous risk to experience an event – the hazard λ_1 in the treatment group and λ_0 in the control group – at each event time $t\langle i\rangle$ (Klein and Moeschberger, 2006, p. 206):

$$\begin{aligned} H_0: &\quad \lambda_1(t\langle i\rangle) = \lambda_0(t\langle i\rangle), && \text{for all event times } i,\\ H_1: &\quad \lambda_1(t\langle i\rangle) \text{ and } \lambda_0(t\langle i\rangle) \text{ are different}, && \text{for some event times } i. \end{aligned}$$

Logrank statistic for single events Let $Y_1\langle i\rangle$ denote that number of participants in the risk set that are in the treatment group at the time of the i^{th} event and $Y_0\langle i\rangle$ for the number of participants at risk in the control group. We use the standard notational convention that $Y_1\langle i\rangle + Y_0\langle i\rangle$ includes the participant experiencing the i^{th} event. Let $O_1\langle i\rangle$ and $O_0\langle i\rangle$ count the number of observed events in the treatment group and control group at the i^{th} event time. These always describe a single event in this section: $O_1\langle i\rangle = 1$ and $O_0\langle i\rangle = 0$ if the event occurred in the treatment group, and $O_1\langle i\rangle = 0$ and $O_0\langle i\rangle = 1$ if a single event occurred in the control group. We extend this Bernoulli case to multiple simultaneous events (ties) – in which case $O_1\langle i\rangle$ can be larger than 1 – in Section 2.1.2. The logrank statistic for a sample size of n single events with nonempty risk sets (i.e.

$Y_1\langle i\rangle > 0, Y_0\langle i\rangle > 0$) is the following (for the treatment group 1):

$$Z = \frac{\sum_{i=1}^{n}\{O_1\langle i\rangle - E_1\langle i\rangle\}}{\sqrt{\sum_{i=1}^{n} V_1\langle i\rangle}} \quad \text{with} \quad E_1\langle i\rangle = \frac{Y_1\langle i\rangle}{Y_0\langle i\rangle + Y_1\langle i\rangle}; \quad V_1\langle i\rangle = E_1\langle i\rangle \cdot \big(1 - E_1\langle i\rangle\big). \tag{2.5}$$

For sufficiently large sample size (number of events n), the logrank Z-statistic has an approximate standard normal (Gaussian) distribution under the null hypothesis.

2.1.1 The safe logrank test for single events

Logrank partial likelihood for single events If we assume that our hazards follow the Cox (1972) proportional hazards model, we obtain a more explicit alternative hypothesis for the logrank test in terms of a constant hazard ratio θ:

$$\begin{aligned} H_0: &\quad \lambda_1(t\langle i\rangle) = \theta \cdot \lambda_0(t\langle i\rangle), && \text{for all event times } t\langle i\rangle, \text{ with } \theta = 1, \\ H_1: &\quad \lambda_1(t\langle i\rangle) = \theta \cdot \lambda_0(t\langle i\rangle), && \text{for all event times } t\langle i\rangle, \text{ with } \theta \neq 1. \end{aligned}$$

This turns the nonparametric logrank test into a semi-parametric test for which we can formulate a partial likelihood. For now, we model the data by a simplified process in which this partial likelihood is simply a standard likelihood, returning to the 'partial' interpretation further below.

We define a risk set process $\vec{Y}\langle 1\rangle, \vec{Y}\langle 2\rangle, \vec{Y}\langle 3\rangle, \ldots$ with $\vec{Y}\langle i\rangle = \big(Y_0\langle i\rangle, Y_1\langle i\rangle\big)$ that describes how many participants are at risk at each event time in the treatment and control group. For simplicity, we assume no censoring, but the likelihood we develop remains valid if the risk set changes because of left-truncation or noninformative right-censoring. At the time of the first event, everyone is at risk, captured by $Y_1\langle 1\rangle = m_1$, the number of participants allocated to the treatment group, and $Y_0\langle 1\rangle = m_0$, the number allocated to the control group. At each event time i, if $O_1\langle i\rangle = 1$ (event in treatment group) then $Y_1\langle i+1\rangle := Y_1\langle i\rangle - 1$ and $Y_0\langle i+1\rangle = Y_0\langle i\rangle$, and analogously for the control group. To complete the specification of the process, we now specify the probability that the single event occurs in the treatment group, given that we have reached a new event time i. This probability only depends on the risk set at the i^{th} event time, so we set:

$$\begin{aligned} &P_\theta(O_1\langle i\rangle = o_1 \mid \vec{Y}\langle 1\rangle, \vec{Y}\langle 2\rangle, \ldots, \vec{Y}\langle i-1\rangle, \vec{Y}\langle i\rangle) := P_\theta(O_1\langle i\rangle = o_1 \mid \vec{Y}\langle i\rangle); \\ &P_\theta(O_1\langle i\rangle = o_1 \mid \vec{Y}\langle i\rangle = (y_0, y_1)) := q_\theta(o_1 \mid (y_0, y_1)), \end{aligned}$$

where $o_1 \in \{0, 1\}$ and q_θ is the conditional probability mass function of a treatment group event, given that the i^{th} event occurred. That is,

$$q_\theta\big(o_1 \mid (y_1, y_0)\big) = \left(\frac{y_1 \cdot \theta}{y_0 + y_1 \cdot \theta}\right)^{o_1} \left(\frac{y_0}{y_0 + y_1 \cdot \theta}\right)^{1-o_1} \tag{2.6}$$

is the probability mass function of a Bernoulli $y_1\theta/(y_0 + y_1\theta)$-distribution. This really expresses that, given that there is an event, the probability of being the participant with this event is $\theta/(y_0 + y_1\theta)$ for each participant in the treatment group, and $1/(y_0 + y_1\theta)$ for each participant in the control group. Summing the probabilities gives (2.6).

Our process is fully specified by the product of conditional probability mass functions (2.6). In particular, at the sample size of n observed events, the joint likelihood (probability mass of these n events as a function of the parameter θ conditional on the data) is given by

$$\mathfrak{L}\big(\theta \mid \vec{Y}\langle 1\rangle, \vec{Y}\langle 2\rangle, \ldots, \vec{Y}\langle n\rangle\big) = \prod_{i=1}^{n} q_\theta\big(O_1\langle i\rangle \mid (Y_0\langle i\rangle, Y_1\langle i\rangle)\big). \tag{2.7}$$

We can think of (2.7) as a standard likelihood in the sample space implicitly defined above, giving an extremely simplified handling of time. As we shall see in Section 2.4.3, we can also think of it as a special case of Cox' partial likelihood, obtained if there are no covariates except for the treatment/control indicator. Further below we shall motivate it further in terms of continuous time. Below, we illustrate the connection to the classical logrank test.

Logrank score test for single events If we define $\beta = \log\theta$ (following Cox (1972)), take the score function $U(\beta)$ of the likelihood in (2.7) and evaluate at $\beta = 0$, we get the familiar ingredients of logrank statistic in terms of observed and expected events:

$$U(\beta) = \frac{\mathrm{d}\log \mathfrak{L}(\beta \mid \vec{Y}\langle 1\rangle, \vec{Y}\langle 2\rangle, \ldots, \vec{Y}\langle n\rangle)}{\mathrm{d}\beta} = \sum_{i=1}^{n}\left\{O_1\langle i\rangle - \frac{Y_1\langle i\rangle \cdot \exp(\beta)}{Y_0\langle i\rangle + Y_1\langle i\rangle \cdot \exp(\beta)}\right\},$$

$$-U'(\beta) = -\frac{\mathrm{d}^2 \log \mathfrak{L}(\beta \mid \vec{Y}\langle 1\rangle, \ldots, \vec{Y}\langle n\rangle)}{\mathrm{d}\beta^2} = \sum_{i=1}^{n}\left\{\frac{Y_1\langle i\rangle \exp(\beta)}{Y_0\langle i\rangle + Y_1\langle i\rangle \exp(\beta)}\frac{Y_0\langle i\rangle}{Y_0\langle i\rangle + Y_1\langle i\rangle \exp(\beta)}\right\},$$

$$U(0) = \sum_{i=1}^{n}\left\{O_1\langle i\rangle - \frac{Y_1\langle i\rangle}{Y_0\langle i\rangle + Y_1\langle i\rangle}\right\} = \sum_{i=1}^{n}\big\{O_1\langle i\rangle - E_1\langle i\rangle\big\},\ -U'(0) = \sum_{i=1}^{n} E_1\langle i\rangle\big(1 - E_1\langle i\rangle\big).$$

Under the null hypothesis, standardizing $U(0)$ with the variance $-U'(0)$ (dividing by the observed Fisher information) gives the logrank Z-statistic in (2.5). This shows that the logrank test is the score test corresponding to the likelihood (2.7), a fact already expressed by Cox (1972) when introducing the proportional hazards model. A more detailed derivation is provided in Appendix Section 2.C.

Logrank E-variable For given $\theta_0, \theta_1 > 0$, define the following one-outcome likelihood ratio

$$M_{\theta_1,\theta_0}\langle i\rangle := \frac{\mathfrak{L}\big(\theta_1 \mid \vec{Y}\langle i\rangle\big)}{\mathfrak{L}\big(\theta_0 \mid \vec{Y}\langle i\rangle\big)} = \frac{q_{\theta_1}(O_1\langle i\rangle \mid Y_1\langle i\rangle, Y_0\langle i\rangle)}{q_{\theta_0}(O_1\langle i\rangle \mid Y_1\langle i\rangle, Y_0\langle i\rangle)}. \tag{2.8}$$

Since the likelihood ratio of the i^{th} event time depends only on the risk set at the i^{th} event time $\vec{Y}\langle i\rangle$, we can write out the expectation as follows

$$\begin{aligned}\mathbf{E}_{P_{\theta_0}}\big[M_{\theta_1,\theta_0}\langle i\rangle \mid \vec{Y}\langle 1\rangle, \ldots, \vec{Y}\langle i\rangle\big] &= \mathbf{E}_{P_{\theta_0}}\big[M_{\theta_1,\theta_0}\langle i\rangle \mid \vec{Y}\langle i\rangle = (y_1, y_0)\big] \\ &= \sum_{o_1\in\{0,1\}} q_{\theta_0}(o_1 \mid y_1, y_0)\cdot\frac{q_{\theta_1}(o_1 \mid y_1, y_0)}{q_{\theta_0}(o_1 \mid y_1, y_0)} = \sum_{o_1\in\{0,1\}} q_{\theta_1}(o_1 \mid y_1, y_0) = 1.\end{aligned} \tag{2.9}$$

This standard argument (as in Example 1) immediately shows that, under P_{θ_0}, for all i and all $\theta_1 > 0$, $M_{\theta_1,\theta_0}\langle i\rangle$ is an E-variable conditional on $\vec{Y}\langle 1\rangle, \ldots, \vec{Y}\langle i\rangle$, and $M^{\langle 1\rangle}_{\theta_1,\theta_0}, M^{\langle 2\rangle}_{\theta_1,\theta_0}, \ldots$ with

$$M^{\langle n\rangle}_{\theta_1,\theta_0} := \prod_{i=1}^{n} M_{\theta_1,\theta_0}\langle i\rangle \tag{2.10}$$

is a test martingale under P_{θ_0} relative to the risk set process $\vec{Y}\langle 1\rangle, \vec{Y}\langle 2\rangle, \vec{Y}\langle 3\rangle, \ldots$. Here we note that $M_{\theta_1,\theta_0}\langle i\rangle$ corresponds to $S\langle i+1\rangle$ in Definition 2.0.2, and similarly $M^{\langle i\rangle}_{\theta_1,\theta_0}$ corresponds to $S^{\langle i+1\rangle}$, reflecting existing different notational conventions in the test martingale and survival analysis literature.

Thus, by Ville's inequality, we have the desired:

$$P_{\theta_0}\Big(\text{there exists } n \text{ with } M^{\langle n\rangle}_{\theta_1,\theta_0} \geq 1/\alpha\Big) \leq \alpha. \tag{2.11}$$

We can generalize M_{θ_1,θ_0} by replacing q_{θ_1} in (2.8) by another conditional probability mass function $r_i(\cdot \mid \ldots)$ on $x \in \{0,1\}$, allowed to depend on i and the full past sequence of risk sets $Y\langle 1\rangle, \ldots, Y\langle i\rangle$. For any given sequence of such conditional probability mass functions, $r \equiv \{r_i\}_{i\in\mathbb{N}}$, we extend definition (2.8) to

$$M_{r,\theta_0}\langle i\rangle = \frac{r_i(O_1\langle i\rangle \mid \vec{Y}\langle 1\rangle, \ldots, \vec{Y}\langle i\rangle)}{q_{\theta_0}(O_1\langle i\rangle \mid Y_1\langle i\rangle, Y_0\langle i\rangle)} \quad ; \quad M^{\langle n\rangle}_{r,\theta_0} := \prod_{i=1}^{n} M_{r,\theta_0}\langle i\rangle. \tag{2.12}$$

For any choice of the r_i, the analogue of (2.9) clearly still holds for the resulting M_{r,θ_0}, making $M_{r,\theta_0}\langle i\rangle$ a conditional E-variable and its product process a martingale; and then Ville's inequality (2.11) must also still hold.

The Underlying Continuous Time Model The logrank test is usually justified based on an underlying continuous time model. We now show that tests using $M^{\langle n\rangle}_{r,\theta_0}$ remain valid under this same underlying model, which we now describe. We identify each of the $m = m_0 + m_1$ participants by an index $j \in \{1, \ldots, m\}$. The vector $\vec{g} \in \{0,1\}^m$ encodes group assignment, $\vec{g}_j$ denoting the group assignment of participant j, so that the number of 0-components of $\vec{g}$ is m_0. We assume that the group assignment is fixed in advance. For simplicity we assume no censoring, but as said, all results remain valid if the risk set changes because of left-truncation or noninformative right-censoring. We assume that each participant $j \in \{1, \ldots, m\}$ has an event at time T_j, where $T_1, T_2, \ldots, T_m$ are independent, and T_j has distribution that is fully determined by j's group membership: T_j has distribution $P_{\vec{g}_j}$ if patient j is in group g. For each participant j we let $S_j(t) = P_{\vec{g}_j}\{T_j > t\}$ be its survival function and let $\lambda_{\vec{g}_j}(t) = -\frac{\mathrm{d}}{\mathrm{d}t}\ln S_i(t)$ its hazard function, which we assume exists and is continuous. It will then depend on the participant only through its group membership. The relation between the survival function and the hazard function implies that the conditional probability that the event time T_j falls in the interval $(t, t+h]$ given that $T_j > t$, can be computed as

$$P_{\vec{g}_j}\{t < T_j \leq t+h \mid T_j > t\} = 1 - \exp\Big(\int_t^{t+h} \lambda_{\vec{g}_j}(s)\,\mathrm{d}s\Big) \tag{2.13}$$

for any $t > 0$. Under the proportional hazards ratio model, the ratio $\lambda_1(t)/\lambda_0(t)$ remains constant and takes the value θ. The results of Slud (1992) imply that under these assumptions (i.e. independence of T_j and proportional hazards), the conditional distribution of $O_1\langle i\rangle$ given $Y_1\langle i\rangle, Y_0\langle i\rangle$ is indeed uniquely defined and given by $q_{\theta_0}(O_1\langle i\rangle \mid Y_1\langle i\rangle, Y_0\langle i\rangle)$ (see also Andersen et al. (1993) for closely related results); for convenience we give an informal derivation (avoiding measure theory and ignoring censoring) of this result in Appendix Section 2.A.

Combining this result with (2.2) we see that $M^{\langle n\rangle}_{r,\theta_0}$ remains a test martingale within this standard continuous time model, and Ville's inequality remains valid.

Example 2. [GROW alternative] The simplest possible scenario is that of a one-sided test between the 'no effect' null hypothesis H_0 ($\theta_0 = 1$) and a one-sided alternative hypothesis $H_1 = \{P_{\theta_1} : \theta_1 \in \Theta_1\}$ represented by a minimal clinically relevant effect size $\theta_{\min}$. For example, if 'event' means that the participant gets ill, then we would hope that under the treatment, $\theta_{\min}$ would be a value smaller than 1 and we would have $\Theta_1 = \{\theta_1 : 0 < \theta_1 \leq \theta_{\min}\}$. If 'event' means 'cured' then we would typically set $\Theta_1 = \{\theta_1 : \theta_{\min} \leq \theta_1 < \infty\}$ for some $\theta_{\min} > 1$. We will take the left-sided alternative with $\theta_{\min} < 1$ as a running example, but everything we say in the remainder of this chapter also holds for the right-sided alternative. The GROW (*growth-optimal in worst-case*) E-variable as in (2.4) is given by taking $M_{\theta_{\min},\theta_0}$, i.e. it takes the $\theta_1 \in \Theta_1$ closest to θ_0. That is,

$$\max_{\theta_1>0} \min_{\theta\in\Theta_1} \mathbf{E}_{P_\theta}[\log M_{\theta_1,\theta_0}\langle i\rangle \mid Y_1\langle i\rangle, Y_0\langle i\rangle] = \max_{r} \min_{\theta\in\Theta_1} \mathbf{E}_{P_\theta}[\log M_{r,\theta_0}\langle i\rangle \mid Y_1\langle i\rangle, Y_0\langle i\rangle]$$

is achieved by setting $\theta_1 = \theta_{\min}$, no matter the values taken by $Y_1\langle i\rangle, Y_0\langle i\rangle$. Here the second maximum is over all sequences of conditional distributions r_i as used in (2.12). Thus, among all E-variables of the general form $M_{r,\theta_0}\langle i\rangle$ there are strong reasons for setting $r_i = q_{\theta_{\min}}$ – this is further elaborated in Section 2.4 and Appendix Section 2.B.

Now suppose we want to do a two-sided test, with alternative hypothesis $\{P_{\theta_1} : \theta_1 \leq \theta_{\min} \vee \theta_1 \geq 1/\theta_{\min}\}$ with $\theta_{\min} < 1$. For this case, one can create a new 'combined GROW' E-variable

$$M^{\langle i\rangle}_{\text{two-sided}} := \frac{1}{2}\left(M^{\langle i\rangle}_{\theta_{\min},\theta_0} + M^{\langle i\rangle}_{1/\theta_{\min},\theta_0}\right). \tag{2.14}$$

It is then easy to verify that for each i, $M_{\text{two-sided}}\langle i\rangle := M^{\langle i\rangle}_{\text{two-sided}}/M^{\langle i-1\rangle}_{\text{two-sided}}$ is a conditional E-variable, i.e. $\mathbf{E}_{P_{\theta_0}}\left[M_{\theta_1,\theta_0}\langle i\rangle \mid Y_1\langle i\rangle, Y_0\langle i\rangle\right] = 1$, and $M_{\text{two-sided}}\langle 1\rangle, M_{\text{two-sided}}\langle 2\rangle, \ldots$ constitutes a test martingale; see GHK for details.

Example 3. [Meta-analysis] What if we want to combine several trials, conducted in different hospitals or in different countries? In such a case we often compare a 'global' null – H_0 is true in all trials – to an alternative that allows for different hazard ratios in different trials, with different populations. We may thus associate the i^{th} event at the k^{th} trial with E-variable $M_{\theta_{1,k},\theta_0}\langle i,k\rangle$, with $\theta_{1,k}$ varying from trial to trial.

$$M\langle i,k\rangle := \frac{q_{\theta_{1,k}}(O_1\langle i,k\rangle \mid Y_1\langle i,k\rangle, Y_0\langle i,k\rangle)}{q_{\theta_0}(O_1\langle i,k\rangle \mid Y_1\langle i,k\rangle, Y_0\langle i,k\rangle)}$$

denotes the E-variable corresponding to the i^{th} event in the k^{th} trial, with $Y_1\langle i,k\rangle$ and $Y_0\langle i,k\rangle$ denoting the number of people at risk in the treatment and control group of trial k at the i^{th} event. The evidence against H_0 after having observed n_k events from trial k, for $k \in \mathscr{K}$, with $\mathscr{K}$ the subset of all trials for which some data is already available can then be summarized as

$$M_{\text{META}} := \prod_{k\in\mathscr{K}} \prod_{i=1,\ldots,n_k} M\langle i,k\rangle.$$

As GHK explain, in such cases the anytime-valid type-I error guarantee still holds: under the global null, where $\theta = \theta_0$ in all trials, the probability that there ever comes a sequence of events in any combination of trials such that for this sequence, $M_{\text{META}} \geq 1/\alpha$, is still bounded by α. Thus, we effectively perform an *on-line, cumulative* and possibly *live* meta-analysis here that remains valid irrespective of the order in which the events of the different trials come in. Importantly, unlike in α-spending approaches, the maximum number of trials and the maximum sample size (number of events) per trial do not have to be fixed in advance; we can always decide to start a new trial, or to postpone to end a trial and wait for additional events. This has many advantages in terms of collaboartion, efficiency and communication of results (see Chapter 1).

2.1.2 Allowing for ties

In many settings we may observe ties: we cannot identify distinct event times for all observed events and consider some of those events to have happened simultaneously. To formalize this we first define $Y\langle i\rangle = Y_1\langle i\rangle + Y_0\langle i\rangle$ and $O\langle i\rangle = O_1\langle i\rangle + O_0\langle i\rangle$. In case of a tie at event time i, $O\langle i\rangle$ is larger than one and consists of multiple events in the treatment group ($O_1\langle i\rangle \geq 1$) and/or in the control group ($O_0\langle i\rangle \geq 1$). We cannot represent this common situation with our simple process P_θ from Section 2.1.1, since it requires a fully observable ordering of events. But we can easily extend it to ties by conditioning on the number of observed events $O\langle i\rangle = o$ at each of the event times i with multiple events. The probability distribution of the number of events in the treatment group $O_1\langle i\rangle$ is defined by the Fisher noncentral hypergeometric distribution. Here $O_1\langle i\rangle = o_1$ indicates the number of events that are observed in the treatment group, out of a total of $O\langle i\rangle = o$ events (so $O_1\langle i\rangle$ can be $0, 1, \ldots, o$). The distribution q_θ defined below replaces q_θ from (2.7):

$$q_\theta(o_1 \mid (y_0,y_1),o) := \frac{\binom{y_1}{o_1}\cdot\binom{y_0}{o-o_1}\cdot\theta^{o_1}}{\sum_{u=o_1^{\min}}^{o_1^{\max}}\binom{y_1}{u}\cdot\binom{y_0}{o-u}\cdot\theta^{u}} \quad \text{with} \quad \begin{cases} o_1^{\min} = \max\{0, o-y_0\} \\ o_1^{\max} = \min\{o, y_1\}. \end{cases} \tag{2.15}$$

This is the probability mass function of a Fisher noncentral hypergeometric distribution with parameters $(O\langle i\rangle, Y_1\langle i\rangle, Y\langle i\rangle, \theta) = (o, y_1, y, \theta)$. In the Appendix Section 2.C we show that also for this partial likelihood, the score test equals the logrank test. When dealing with ties we must distinguish between the number of events n in a sample and the number of event times $I \leq n$ in the same sample. The corresponding martingale $M^{\langle I\rangle}_{\theta_1,\theta_0} = \prod_{i=1}^{I} M_{\theta_1,\theta_0}\langle i\rangle$ is now a product of I E-variables, together covering n events. In Appendix Section 2.A we show that, in the continuous-time model, $M^{\langle I\rangle}_{\theta_1,\theta_0}\langle i\rangle$ is still a conditional E-variable, and $M^{\langle I\rangle}_{\theta_1,\theta_0}$ is a test martingale *if the null* $\theta_0 = 1$, i.e. under the

assumption that there are no differences between the two groups; for other θ_0, it is only an 'approximate' E-variable, becoming more exact the closer the time points at which we check whether event(s) have happened lie together in continuous time. Thus, we have a weaker result than for the case without ties, for which M_{θ_1,θ_0} is a test martingale in the continuous time setting under arbitrary θ_0 — but as long as we test with null $\theta_0 = 1$, all our results are still exact. Again, for simplicity we ignore censoring in the appendix. The present approach for ties can still be adapted to noninformative right censoring under the additional common assumption that the events reported at each observation time precede any censorings, so that censored patients contribute fully to the risk sets under consideration.

Compatibility between q_θ in (2.15) and (2.6) Reassuringly, if at all event times $i = 1, 2, \ldots, I$ we observe only a single event, q_θ from (2.15) and q_θ from (2.7) are the same, since for a single event in the treatment group ($o = 1, o_1 = 1$) and a given $y_1, y_0 > 1$, we get

$$q_\theta(1 \mid (y_0, y_1), 1) = \frac{\binom{y_1}{1} \cdot \binom{y_0}{0} \cdot \theta^1}{\sum_{v=0}^{1} \binom{y_1}{u} \cdot \binom{y_0}{1-u} \cdot \theta^u} = \frac{y_1 \cdot \theta}{y_0 + y_1 \cdot \theta},$$

and analogously for a single event in the control group ($o = 1, o_1 = 0$).

2.1.3 Gaussian approximation on the logrank statistic

Our GROW safe logrank test is an exact logrank test for a risk set process with a hypergeometric probability mass function (2.15) under the null hypothesis ($\theta = 1$) – with the Bernoulli probability mass function (2.6) as a special case for single events. The exact test can be used for a survival data set of event and censoring times that captures the full risk set process. Here we describe an approximation to this safe logrank test based on a sequential-Gaussian approximation on the logrank statistic. The approximation is of interest for two reasons. First, in practical situations, sometimes only the logrank Z-statistic and other summary statistics are available, and not the full risk set process. If we also know the number of events n and initial number of participants in the two groups m_1 and m_0, the Gaussian approximation can then still be used. Second, the group sequential and α-spending approaches that we compare ourselves to in the next section are based on a Gaussian approximation to the logrank statistic. The behavior of the Gaussian approximation can (and will) give us insights into how the safe logrank test compares to the group sequential and α-spending approaches as well.

For a mix of single and tied events, the logrank Z-statistic for n events at I event times is defined as follows:

$$Z = \frac{\sum_{i=1}^{I} \{O_1\langle i\rangle - E_1\langle i\rangle\}}{\sqrt{\sum_{i=1}^{I} V_1\langle i\rangle}} \quad \text{where} \tag{2.16}$$

$$E_1\langle i\rangle = O\langle i\rangle \cdot A_1\langle i\rangle; \qquad A_1\langle i\rangle = \frac{Y_1\langle i\rangle}{Y\langle i\rangle}; \qquad V_1\langle i\rangle = O\langle i\rangle \cdot A_1\langle i\rangle \cdot \big(1 - A_1\langle i\rangle\big) \cdot \frac{Y\langle i\rangle - O\langle i\rangle}{Y\langle i\rangle - 1}.$$

For single events, $O\langle i\rangle = 1$ and $E_1\langle i\rangle = A_1\langle i\rangle$. If all event times have single events, then $n = I$. In general, with and without ties, $n = \sum_{i=1}^{I} O\langle i\rangle$. The above formulation is also found in Cox (1972, equation (26)) and $\frac{Y\langle i\rangle - O\langle i\rangle}{Y\langle i\rangle - 1}$ is described as a "multiplicity" correction or "correction for ties" (Klein and Moeschberger, 2006, p. 207).

Under the null hypothesis, this logrank statistic has an approximate standard normal (Gaussian) distribution ($Z \sim \mathcal{N}(0,1)$). Under the alternative distribution, Schoenfeld (1981) gives an asymptotic result for survival data (no ties) following the proportional hazards model with hazard ratio θ: a Taylor approximation around $\theta = 1$ gives that the logrank Z-statistic is also normally distributed ($Z \sim \mathcal{N}(\mu, 1)$) with:

$$\mu \approx \frac{\sum_{i=1}^{n} \log(\theta) E_1\langle i\rangle \big(1 - E_1\langle i\rangle\big)}{\sqrt{\sum_{i=1}^{n} E_1\langle i\rangle \big(1 - E_1\langle i\rangle\big)}} \approx \log(\theta)\sqrt{nE_1\langle 1\rangle\big(1 - E_1\langle 1\rangle\big)} = \log(\theta)\sqrt{\frac{m_1 \cdot m_0}{(m_1 + m_0)^2}}\sqrt{n}. \tag{2.17}$$

Schoenfeld's assumptions Schoenfeld's asymptotic result heavily relies on two properties: (a) the mean of the alternative is close enough to one so that the first-order Taylor approximation around $\theta = 1$ is adequate and (b) $E_1\langle i\rangle$ stays approximately constant at all event times i, i.e. close to the initial allocation proportion $E_1\langle 1\rangle = m_1/(m_1 + m_0)$. These two properties indicate that this asymptotic distribution is only reasonably good if the hazard ratio θ is close to 1 and the initial risk set m_0 and m_1 are both large in comparison to the number of events and the amount of censoring, in which case also the multiplicity correction in the definition of $V_1\langle i\rangle$ for ties is negligible.

Logrank statistic per event time This raises the question whether a Gaussian approximation is sensible for a logrank statistic per event time i: a priori it is not at all clear whether Schoenfeld's asymptotic, fixed sample result has a nonasymptotic sequential counterpart. We define the logrank statistic per event time

$$Z\langle i\rangle = \frac{O_1\langle i\rangle - E_1\langle i\rangle}{\sqrt{V_1\langle i\rangle}}. \tag{2.18}$$

and check whether the exact E-variable (with the q for a mix of ties and single events from (2.15))

$$M_{\theta_1,\theta_0}\langle i\rangle = \frac{q_{\theta_1}\big(O_1\langle i\rangle \mid (Y_1\langle i\rangle, Y_0\langle i\rangle), O\langle i\rangle\big)}{q_{\theta_0}\big(O_1\langle i\rangle \mid (Y_1\langle i\rangle, Y_0\langle i\rangle), O\langle i\rangle\big)} \text{ behaves similar to } M'_{\mu_1,\mu_0}\langle i\rangle := \frac{\phi_{\mu_1\sqrt{O\langle i\rangle}}(Z\langle i\rangle)}{\phi_{\mu_0}(Z\langle i\rangle)} \tag{2.19}$$

for $\theta_0 = 1, \mu_0 = 0$ and $\mu_1 = \log(\theta_1) \cdot \sqrt{(m_1 \cdot m_0)/(m_1 + m_0)^2}$, where ϕ_μ is the Gaussian density with mean μ and variance 1.

Safety only for balanced allocation We henceforth focus on the case $\theta_0 = 1$, such that that $\mu_0 = 0$. Figure 2.1 shows that in case of balanced 1:1 allocation $M'_{\mu_1,0}\langle i\rangle$ is an E-variable, since its expectation is 1 or smaller. However, in case of unbalanced 2:1 or

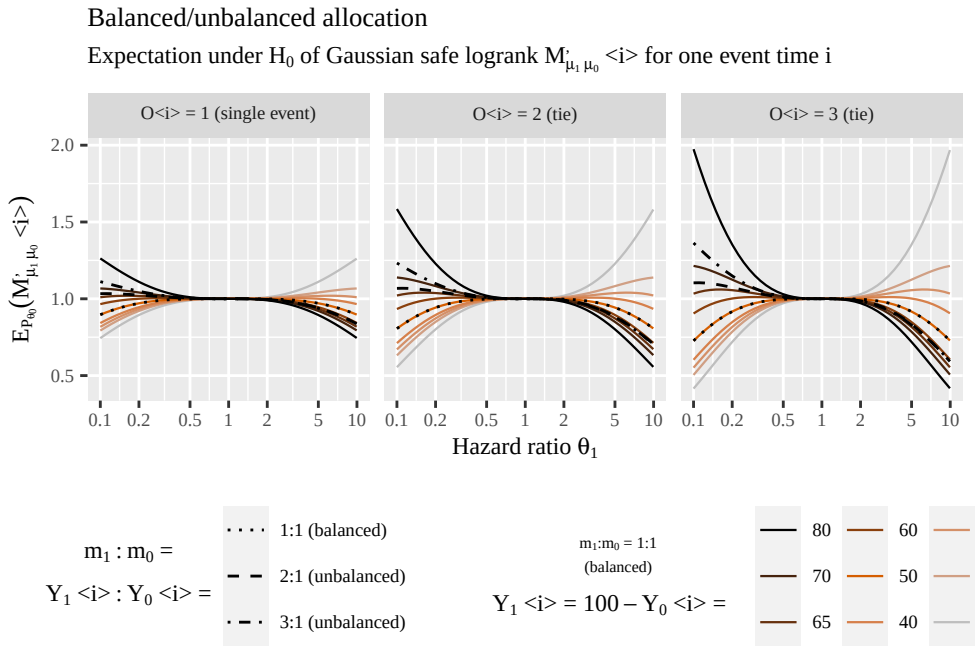

Figure 2.1. *For balanced allocation $M'_{\mu_1,\mu_0}\langle i\rangle$ is an E-variable, but it is not for unbalanced allocation. The risk set can also start out balanced but become unbalanced, in which case $M'_{\mu_1,\mu_0}\langle i\rangle$ is also not an E-variable. Here $\mu_0 = 0$, and μ_1 follows from θ_1 as in (2.19); in the expectation $E_{P_{\theta_0}}$ under the null hypothesis $\theta_0 = 1$. Note that for $i > 1$, $M'_{\mu_1,\mu_0}\langle i\rangle$ depends both on the present risk set balance $Y_1\langle i\rangle : Y_0\langle i\rangle$ and on the initial balance $m_1 : m_0$. Note also that the x-axis is logarithmic.*

3:1 allocation and designs with hazard ratio $\theta_1 < 1$, $M'_{\mu_1,0}\langle i\rangle$ is not an E-variable. Of course, the risk set can also start out balanced by allocation, but become unbalanced. Figure 2.1 shows that in case of designs outside the range $0.5 \leq \theta_1 \leq 2$ the deviations from expectation 1 can be problematic. Hence we recommend to not use the Gaussian approximation on the logrank statistic for unbalanced designs and designs for $\theta_1 < 0.5$ or $\theta_1 > 2$. For balanced designs with $0.5 \leq \theta_1 \leq 2$, we found that in practice they are safe to use as long as the risk set is large in comparison to the number of events and the amount of censoring, the reason being that scenarios in which the allocation becomes highly unbalanced after some time (e.g. $Y_1\langle i\rangle = 80, Y_0\langle i\rangle = 20$) are extremely unlikely under the null.

Optimality close to hazard ratio 1 In case of balanced allocation, Figure 2.2 shows that the approximate e-values for a single event time from the Gaussian $M'_{\mu_1,0}\langle i\rangle$ are very similar to the exact e-values $M_{\theta_1,1}\langle i\rangle$ in designs for alternative hazard ratios θ_1 between

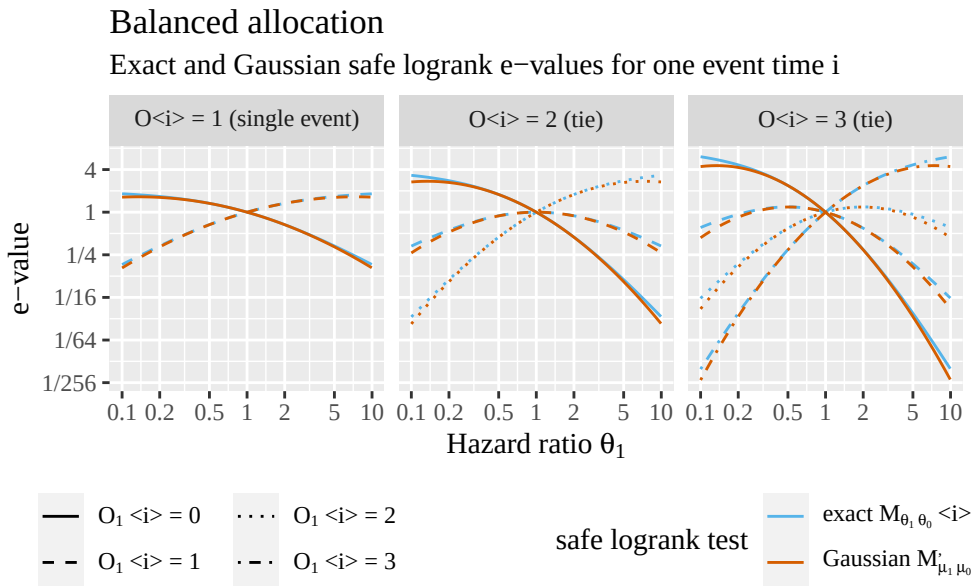

Figure 2.2. For balanced allocation ($m_1 = m_0 = Y_1\langle i\rangle = Y_0\langle i\rangle$) $M'_{\mu_1,\mu_0}\langle i\rangle$ is very similar to $M_{\theta_1,\theta_0}\langle i\rangle$ in case of designs for $0.5 \le \theta_1 \le 2$. Here $\theta_0 = 1$, $\mu_0 = 0$, and μ_1 follows from θ_1 as in (2.19). Note that both axes are logarithmic.

0.5 and 2. Having observed I event times, leading to a logrank statistic Z as in (2.16), we can therefore directly approximate $M^{\langle I\rangle}_{\theta_1,1}$ by

$$M''_{\mu_1,0} = \frac{\phi_{\mu_1\sqrt{n}}(Z)}{\phi_0(Z)} \quad \text{with } \mu_1 = \log(\theta_1)\cdot\sqrt{\frac{m_1\cdot m_0}{(m_1+m_0)^2}}. \tag{2.20}$$

This second approximation is valid under Schoenfeld's assumption (b) of constant large risk set. For then the initial risk set sizes m_0 and m_1 are both large in which case also the multiplicity correction in the definition of $V_1\langle i\rangle$ for ties is negligible. Without the multiplicity correction, the variance $V_1\langle i\rangle$ in (2.16) is equal to the variance of $O\langle i\rangle$ single events. Again, if the initial risk sets are sufficiently large, we can treat this variance to be constant in i and equal to $m_1 m_0/(m_1+m_0)^2$ per event. Taken together straightforward

calculus thus gives

$$\begin{aligned}\prod_{i=1}^{I} M'_{\mu_1,0}\langle i\rangle &= \exp\left(-\frac{1}{2}\sum_{i=1}^{I}\left(\mu_1^2 O\langle i\rangle - 2\mu_1\sqrt{O\langle i\rangle}Z\langle i\rangle\right)\right)\\ &\approx \exp\left(-\frac{1}{2}n\mu_1^2 + \mu_1\sum_{i=1}^{I}\sqrt{O\langle i\rangle}\,\frac{O_1\langle i\rangle - E_1\langle i\rangle}{\sqrt{O\langle i\rangle}\sqrt{\frac{m_1\cdot m_0}{(m_1+m_0)^2}}}\right)\\ &= \exp\left(-\frac{1}{2}n\mu_1^2 + \mu_1\frac{\sum_{i=1}^{I}\left\{O_1\langle i\rangle - E_1\langle i\rangle\right\}}{\sqrt{\frac{m_1\cdot m_0}{(m_1+m_0)^2}}}\right)\\ &\approx \exp\left(-\frac{1}{2}n\mu_1^2 + \mu_1\sqrt{n}Z\right) = M''_{\mu_1,0}.\end{aligned}$$

The first $\approx$ uses $\sum_{i=1}^{I} O\langle i\rangle = n$ and approximates the Z-score per event time from (2.18) assuming that the correction for ties is negligible and the risk set constant; the second $\approx$ maps back to the logrank statistic of multiple events Z using that the single event variance $m_1 \cdot m_0/(m_1+m_0)^2$ is approximately n times smaller than the variance $\sum_{i=1}^{I} V_1\langle i\rangle$ used to define Z in (2.16).

We will call $M''_{\mu_1,0}$ the *approximately safe logrank test based on the Gaussian approximation on the logrank statistic* or simply the *Gaussian safe logrank test*. In Figure 2.3 we investigate the power of this Gaussian safe logrank test and confirm that it often achieves e-value > 20 at exactly the same sample size as the exact safe logrank test. (In Section 2.3 we investigate this further.) Thus, a hazard ratio between 0.5 and 1 – and by symmetry also between 1 and 2 – is apparently sufficiently close to 1 (Schoenfeld's assumption (a)). Also, the deviations from the constant large and balanced risk set do not seem to occur often for this range of hazard ratios (Schoenfeld's assumption (b)). After all, the risk set needs to be quite large to observe the number of events to detect hazard ratios in the range $0.5 \leq \theta_1 \leq 2$. Figure 2.2 and Figure 2.3 show that the closer θ_1 is to 1, the more similar the Gaussian safe logrank test and the exact safe logrank test behave, but that the exact safe logrank test is optimal in general.

2.2 Comparing rejection regions

The first approaches to sequential analysis came with precise stopping rules. Wald's Sequential Probability Ratio Test (SPRT) (Wald, 1947), for example, specifies an upper and a lower boundary and requires the experiment to stop when either of the two is crossed. Hence the guarantees for type-I error control in crossing the upper boundary strictly rely on the lower boundary and vice versa.

Sequential analysis became popular in monitoring of clinical trials (by so-called Data Safety and Monitoring Boards, DSMBs) with the arrival of group-sequential methods – especially when timing interim analyses was flexibilized by α-spending functions (Lan

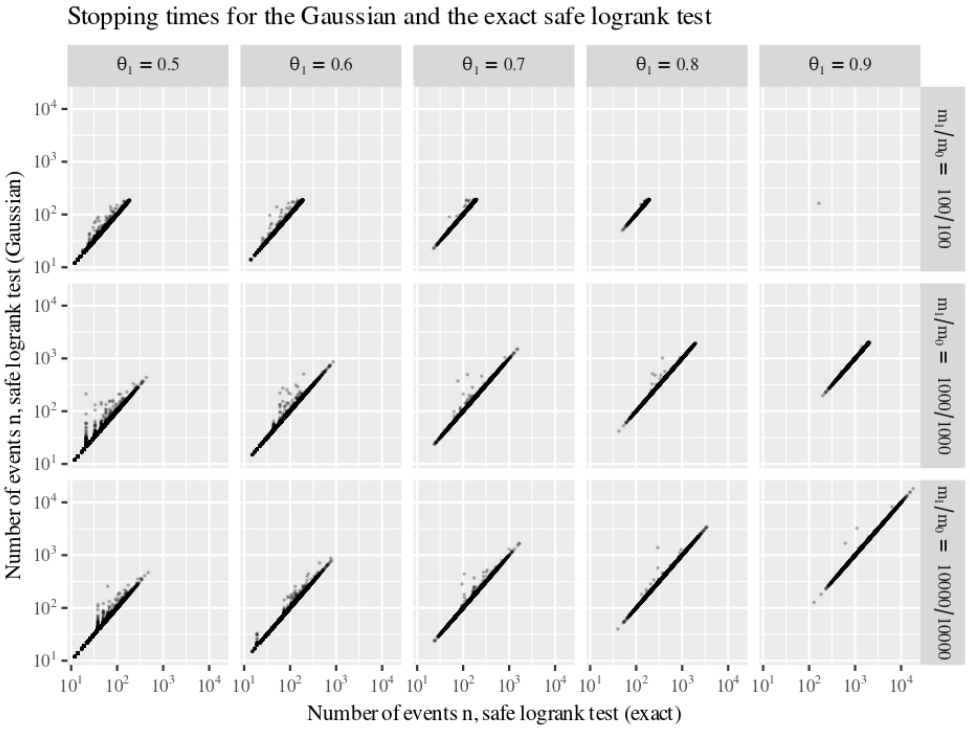

Figure 2.3. *Stopping times $I = n$ (number of events before stopping) simulated as single events to achieve e-value > 20 (safe test with $\alpha = 0.05$), when using the exact logrank martingale $M^{\langle n \rangle}_{\theta_1,1}$ from (2.10) and the Gaussian approximation on the logrank statistic $M''_{\mu_1,0}$ from (2.20). θ_1 specifies both the hazard ratio used for designing the tests and the true hazard ratio used to simulate the risk set process. Details of the simulation are given in Appendix Section 2.D. Note that both axes are logarithmic.*

and DeMets, 1983). Proschan et al. (2006, p. 214) note that "data are often not available in the order in which patients have been accrued", and therefore patient-by-patient or event-by-event analyses are not possible while monitoring the data. Moreover, boundaries are sometimes more guidelines than strict stopping rules: "One can regard a clinical trial that compares a new treatment to placebo or to an old treatment as having one clearly defined upper one-sided boundary – the one whose crossing demonstrates benefit – and a number of less well defined one-sided lower boundaries, the ones whose crossing worries the DSMB." (Proschan et al., 2006, p. 6). So two criteria are critical: (1) the analysis can separate benefit from harm, upper boundaries from lower boundaries – ignoring the crossing of one should not invalidate the inference from crossing the other; (2) crossing a boundary in the past should not impair a DSMB that can only convene for decisions at irregular intervals with chronologically incomplete data. These two criteria are fulfilled by α-spending functions, but in fact also by our safe logrank test: Ville's inequality (2.11)

allows for any stopping rule and while the test can be two-sided (2.14), it does not need to be.

Here we compare the region of logrank statistics for which α-spending approaches and the safe logrank test reject the null hypothesis of no effect (hazard ratio $\theta_0 = 1$). We discuss the two main α-spending functions that are inspired by two group-sequential approaches – Pocock (1977) and O'Brien and Fleming (1979) – although our main focus is on the O'Brien-Fleming approach. The Pocock procedure rejects at equally extreme values for the Z-statistic for small sample sizes as for larger ones, while DSMB's usually do not want to stop a trial very early for efficacy. Pocock himself now believes that his boundary is unsuitable (Pocock, 2006). Moreover, in contrast to the Pocock approach, the O'Brien Fleming α-spending approach can be used to monitor the data after each new event-time. Hence the fair assessment is to compare continuously monitoring the safe logrank test to continuously monitoring using the O'Brien-Fleming α-spending function.

2.2.1 GROW safe logrank (Gaussian) vs α-spending

In the previous section, Figure 2.2 and Figure 2.3 show that, in case of balanced allocation, the Gaussian approximation to the logrank statistic behaves very similar to the exact logrank test for certain designs. This is the case if we design a trial with 10000 participants to detect an effect of minimum clinical relevance of $\theta_1 = 0.7$ using a design that is *growth rate optimal in the worst case* (GROW, see Example 2). We will take this trial design as an example to compare the safe logrank test to α-spending. Since the O'Brien-Fleming rejection region can be uniquely defined in terms of the logrank Z-statistic, we compare it to the Gaussian safe logrank test defined on the same logrank statistic.

Gaussian safe logrank Z-rejection region For the Gaussian approximation we can specify the region of values for the logrank Z-statistic that rejects the null hypothesis when the e-value is larger than $1/\alpha$ as follows: rejection takes place for values of Z such that

$$M''_{\mu_1,\mu_0} = \frac{\phi_{\mu_1\sqrt{n}}(Z)}{\phi_{\mu_0}(Z)} = \frac{\exp[-\frac{1}{2}(Z-\mu_1\sqrt{n})^2]}{\exp[-\frac{1}{2}Z^2]} \geq \frac{1}{\alpha} \text{ with } \mu_0 = 0;\ \mu_1 = \log(\theta_1)\sqrt{\frac{m_1 \cdot m_0}{(m_1+m_0)^2}}$$

such that for $m_1 = m_0$, we reject if:
$$\begin{cases} Z \geq \frac{1}{2}\log(\theta_1)\cdot\frac{1}{2}\cdot\sqrt{n} - \frac{\log(\alpha)}{\log(\theta_1)\cdot\frac{1}{2}\cdot\sqrt{n}} & \text{if } \theta_1 > 1 \\ Z \leq \frac{1}{2}\log(\theta_1)\cdot\frac{1}{2}\cdot\sqrt{n} - \frac{\log(\alpha)}{\log(\theta_1)\cdot\frac{1}{2}\cdot\sqrt{n}} & \text{if } \theta_1 < 1. \end{cases} \tag{2.21}$$

O'Brien-Fleming α-spending Z-rejection region The O'Brien-Fleming α-spending function gives a boundary that is constant in $B(n/n_{\max}) = Z/\sqrt{n/n_{\max}}$, where $n_{\max}$ is a maximum number of events that has to be set in advance. Result 5.1 in Proschan et al. (2006)

gives that if we take $B(n/n_{\max})$ to be continuous Brownian motion:

$$P_{\mu_0}\Big[B(n/n_{\max}) > c \quad \text{for some } n \leq n_{\max}\Big] = 2P_{\mu_0}\Big[B(n_{\max}/n_{\max}) > c\Big] = 2P_{\mu_0}\Big[Z > c\Big] = \alpha$$

$$\text{such that} \quad \begin{cases} Z \geq \frac{\Phi^{-1}(1-\alpha/2)}{\sqrt{n/n_{\max}}} & \text{in case of a right-sided test design} \\ Z \leq \frac{\Phi^{-1}(\alpha/2)}{\sqrt{n/n_{\max}}} & \text{in case of a left-sided test design.} \end{cases} \tag{2.22}$$

Figure 2.4 plots both the Gaussian safe logrank and the O'Brien-Fleming α-spending rejection regions. The two regions of Z-statistic values share an important feature: they are more conservative to reject the null hypothesis at small sample sizes than at larger ones, requiring more extreme values for the Z-statistic. This sets them apart from the Pocock spending function that requires equally extreme values for the Z-statistic at small and large sample size. Figure 2.4 shows the boundary of the Pocock spending function for 10 interims. Note that the definition of the safe logrank test rejection region requires a very explicit value for the effect size $\theta_1 = \theta_{\min}$ of minimum clinical relevance, while that value is implicit in the definition of the α-spending rejection region. To specify an maximum sample size $n_{\max}$ to achieve a certain power, you also assume an effect size of minimal interest. A fixed sample size analysis designed to detect a minimum hazard ratio of 0.7 would need a number of events $n_{\max} = n = 195$ to achieve 80% power if the true hazard ratio is 0.7. A sequential analysis using α-spending needs a bit more, a maximum sample size of $n_{\max} = 205$ events for an O'Brien-Fleming spending function and $n_{\max} = 245$ for a Pocock spending function, when we design for 10 looks. We investigate the number of events needed by the safe logrank test in the next section (note that the test allows for unlimited monitoring, so there is no real equivalent to $n_{\max}$).

The benefit of a sequential approach is that if the data looks better than hazard ratio 0.7 we can detect that with a number of events that is smaller than this maximum sample size. Figure 2.5 illustrates that we benefit because the true hazard ratio could be more extreme than we designed for (e.g. 0.5 instead of 0.7; a larger risk reduction in the treatment group) and the data reflects that. We also benefit from a sequential analysis if the true hazard ratio is 0.7 but by chance the values of our Z-statistics are more extreme than expected. The major difference between α-spending approaches and the safe logrank test is that the safe test does not require to set a maximum sample size. It in fact allows to indefinitely increase the sample size without ever spending all α. While an α-spending approach designed to have 80% power will miss out on rejecting the null hypothesis in 20% of cases (the type-II error) – shown to stay green in Figure 2.5 – the safe logrank test can potentially reject all. In the sequences of 500 events in Figure 2.5, all but one sequence of Z-statistics could be rejected at a larger sample size by the safe logrank test. By increasing the sample size, the safe logrank test can have 100% power if the true hazard ratio is at least as small as the hazard ratio set for minimum clinical relevance in the design of the test. Still, type-I errors are controlled. Figure 2.5 shows two null sequences of Z-statistics with a true hazard ratio of 1 that are rejected by the O'Brien-Fleming α-spending region, but not by the safe logrank test. Here, the safe logrank test is a bit more conservative, since it is saving α for the future.

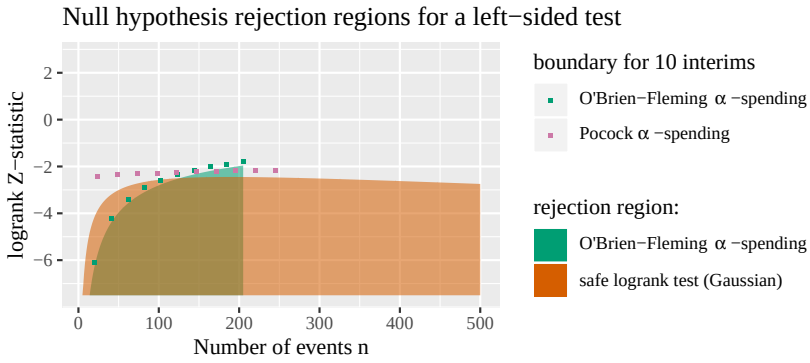

Figure 2.4. H_o rejection regions for continuously monitoring using O'Brien-Fleming α-spending and the safe logrank test, under balanced allocation ($m_0 = m_1$) and for one-sided $\alpha = 0.05$. Also shown are the O'Brien-Fleming and Pocock α-spending boundaries for 10 interim analyses. The α-spending boundaries are designed to have 80% power to detect a hazard ratio 0.7 (left-sided), leading to values of $n_{\max}$ given in the text. The safe logrank test is growth rate optimal for the worst case *hazard ratio $\theta_1 = \theta_{\min} = 0.7$ (left-sided), which we assume is of minimum clinical relevance.*

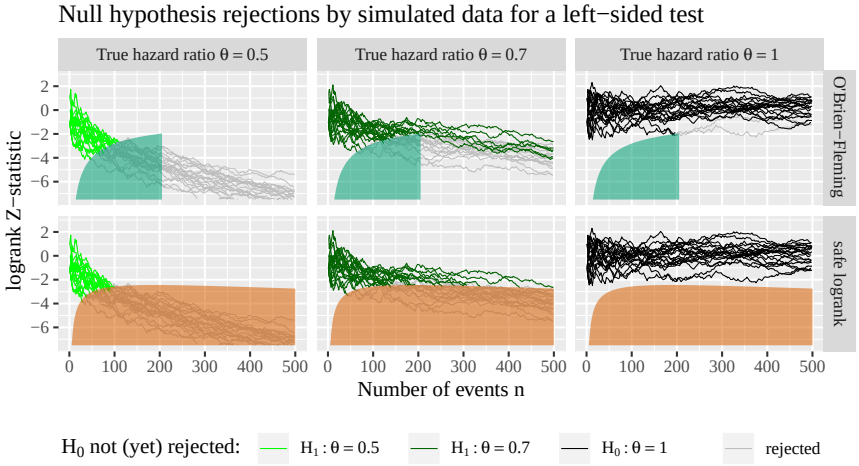

Figure 2.5. H_o rejected by simulated data in rejection regions as described in Figure 2.4 (designed to detect a hazard ratio of 0.7). The data is simulated under balanced allocation ($m_1 = m_0 = 5000$) and as time-to-event data such that ties can occur: the logrank Z-statistic does not have a value for all n; it sometimes jumps with several additional events at a time.

2.2.2 Unbalanced allocation

α-spending methods are known to behave poorly in case of unbalanced allocation (Wu and Xiong, 2017). In Section 2.1.3 we showed that our Gaussian approximation to the logrank test is also not an E-variable in case of unbalanced allocation. Our exact safe logrank test, however, is an E-variable under any allocation since it is defined directly on the risk set process (2.7), that takes into account the allocation. This suggests that in case of unbalanced allocation, the exact logrank test should be preferred over α-spending methods.

2.3 Comparing sample size

Figure 2.6 shows simulation results establishing three types of sample sizes. The leftmost panels ('Maximum') give the sample size required to design a trial. It expresses the maximum number of events $n_{\max}$ that needs to be observed under the alternative to achieve 80% power. In case of the classical logrank test and α-spending designs any further number of events beyond $n_{\max}$ cannot be analyzed. The rightmost panels ('Mean') show the sample size that captures the expected duration of the trial. It expresses the mean number of events, under the alternative, that will be observed before the trial can be stopped (here, for the safe logrank tests, we use the aggressive stopping rule that stops as soon as $M^{\langle n \rangle} \geq 1/\alpha = 20$ or $n = n_{\max}$). In case of α-spending approaches and the safe logrank test this number of events is always smaller than the maximum needed in the design stage. Finally, the middle panel ('Conditional Mean') shows an even smaller number for those tests that have a flexible sample size: the expected stopping time *given* that the trial is stopped before the maximum $n_{\max}$ was reached – which will only happen if the null is rejected. For comparison purposes, all sample sizes are shown relative to (i.e. divided by) the fixed sample size needed by the classical logrank test to obtain 80% power. Note that for small sample size (for small hazard ratios), both the classical logrank test and O'Brien-Fleming α-spending are not recommended due to lack of type-I error control.

GROW safe logrank vs classical logrank and O'Brien-Fleming α-spending Figure 2.6 shows that the classical logrank test and O'Brien-Fleming α-spending require a smaller maximum number of events to obtain 80% power. This is the benefit of specifying a strict $n_{\max}$ in the design and spending all α such that you cannot analyze any further events. The additional number of events required for the maximum of the safe logrank test is the price to pay for the unlimited horizon, or for combining with the data from future trials, after some tial maximum has been reached (optional continuation). The Gaussian and exact safe tests are shown to be similar for $\theta_1 \geq 0.5$, but the Gaussian performs poorly for smaller hazard ratios. In terms of the mean number of events before the trial can be stopped, both the safe logrank test and O'Brien-Fleming outperform the classical standard logrank test at almost all hazard ratios for which the classical logrank test is recommended. The exact safe logrank test always needs more events than O'Brien-Fleming α-spending, but is the only approach that has exact type-I error control with small sample size at the smallest hazard ratios. The conditional mean number of events for the exact logrank test outperforms O'Brien-Fleming for designs with hazard ratio $\theta_1 = 0.1$

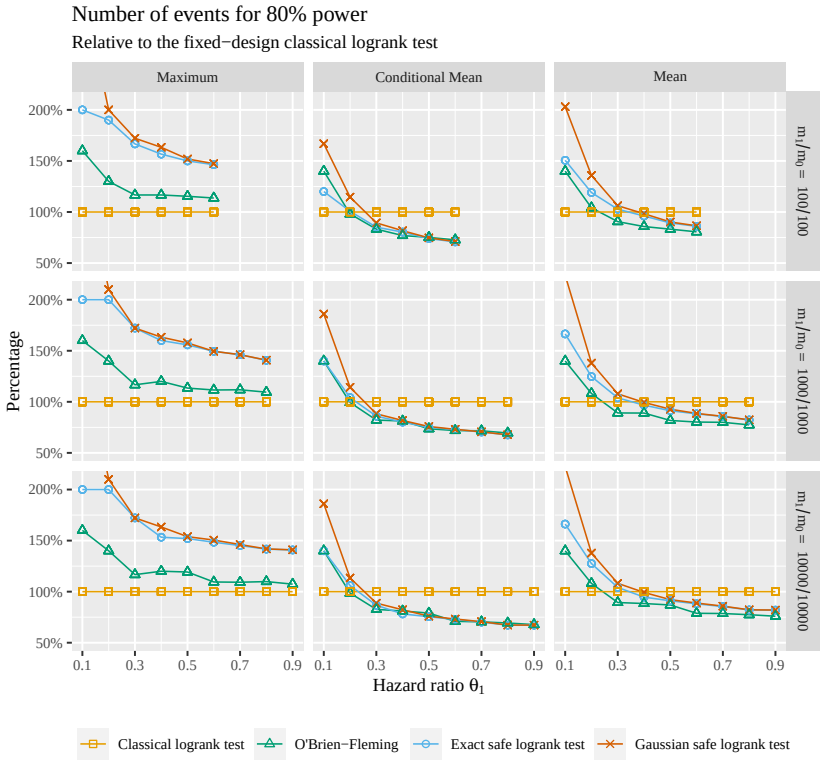

Figure 2.6. *Sample sizes (number of events) needed to reject the null hypothesis with* $\alpha = 0.05$ *using the GROW safe logrank test (exact or Gaussian,* $\theta_0 = 1$, $\mu_0 = 0$*), the classical logrank test (fixed sample size) and O'Brien-Fleming* α*-spending with continuous monitoring (see Section 2.2.1). All tests are designed to detect the hazard ratio* $\theta_1 = \theta_{\min}$ *on the x-axis and data is generated based on that same hazard ratio (see Example 2). The classical logrank test needs the following sample sizes (number of events)* $n(\theta_1)$ *for an* 80% *power design to detect hazard ratio* θ_1*:* $n(0.1) = 5$, $n(0.2) = 10$, $n(0.3) = 18$, $n(0.4) = 30$, $n(0.5) = 52$, $n(0.6) = 95$, $n(0.7) = 195$, $n(0.8) = 497$ *and* $n(0.9) = 2228$ *– these sample sizes represent the* 100% *line in all plots. They are based on Schoenfeld's Gaussian approximation, which underestimates the number of events required for hazard ratios far away from 1 (e.g. simulations show that for* $\theta_1 = 0.1$, $n = 6$ *or 7 events will be necessary) – for small sample sizes the classical logrank test is not recommended due to lack of type-I error control. The difference between Maximum, Conditional Mean and Mean plots is explained in the main text, with further details in Appendix Section 2.D. The upshot is that at all hazard ratios at which the Gaussian approximation to the classical logrank test 'works' (say for* $\theta_1 \geq 0.3$*), the mean number of events needed by the safe logrank tests is about the same or noticeably smaller.*

and small risk set m_1/m_0 and behaves very similar under all other scenarios. While the mean of the stopping times (conditional mean) are similar, the rejection regions in Figure 2.4 illustrate, however, that the stopping times themselves are not the same. The safe logrank test will sometimes reject the null hypothesis at an earlier number of events than O'Brien-Fleming α-spending can, but also stop later than the $n_{\max}$ that was used to design the O'Brien-Fleming α-spending boundary.

2.4 Variations and extensions

2.4.1 Prequential Plug-In and Bayes predictive distributions for the alternative

Now suppose we do not have a very clear idea of which parameter $\theta_1 \in \Theta_1$ to pick. One way to handle this case is to use (2.12) rather than (2.8) and (2.10), replacing the fixed conditional probabilities $q_{\theta_1}(O_1\langle i\rangle \mid Y_1\langle i\rangle, Y_0\langle i\rangle)$ by a conditional probability mass function $r_i(O_1\langle i\rangle \mid \vec{Y}\langle 1\rangle, \ldots, \vec{Y}\langle i\rangle)$ that depends on the past and enables implicitly learning θ_1 from the data. For example, we may take

$$r_i(O_1\langle i\rangle \mid \vec{Y}\langle 1\rangle, \ldots, \vec{Y}\langle i\rangle) := q_{\hat{\theta}_i}(O_1\langle i\rangle \mid Y_1\langle i\rangle, Y_0\langle i\rangle) \tag{2.23}$$

with q as in (2.6) and $\hat{\theta}_i := \hat{\theta}(\vec{Y}\langle 1\rangle, \ldots, \vec{Y}\langle i\rangle)$ the maximum likelihood estimate based on all past data, smoothed by adding two 'virtual' data points to the data, one for both groups:

$$\hat{\theta}(\vec{Y}\langle 1\rangle, \ldots, \vec{Y}\langle i\rangle) :=$$
$$\arg\max_{\theta\in(0,\infty)} \left(\prod_{k=1}^{i} q_\theta(O_1\langle k\rangle \mid Y_1\langle k\rangle, Y_0\langle k\rangle) \right) \cdot q_\theta(1 \mid Y_1\langle 1\rangle{+}1, Y_0\langle 1\rangle) \cdot q_\theta(0 \mid Y_1\langle 1\rangle, Y_0\langle 1\rangle{+}1).$$

Suppose the data are actually sampled from the process defined by q_θ, for some arbitrary but fixed θ. For sufficiently large initial risk sets (m_0, m_1 are not too small), by the law of large numbers, $\hat{\theta}_i$ will converge with high probability to θ, and $q_{\hat{\theta}_i}$ will behave more and more like the real q_θ from which data are sampled. Thus, the process (2.12) instantiated with (2.23) will behave more and more similarly to the 'correct' likelihood ratio (2.10). The process $r_1, r_2, \ldots$ is a typical instance of Dawid's (1984) *prequential plug-in* likelihood, that is often based on suitable smoothed likelihood-based estimators (Grünwald and Roos, 2020). Instead of r_i based on a plug-in estimate of θ based on $\vec{Y}\langle 1\rangle, \ldots, \vec{Y}\langle i\rangle$, one may just as well use a Bayes predictive distribution based on the same data and some prior W_1 on θ. That is, we set

$$r_i(O_1\langle i\rangle \mid \vec{Y}\langle 1\rangle, \ldots, \vec{Y}\langle i\rangle) := q_{W_i}(O_1\langle i\rangle \mid Y_1\langle i\rangle, Y_0\langle i\rangle) \tag{2.24}$$

where $W_i := W \mid \vec{Y}\langle 1\rangle, \ldots, \vec{Y}\langle i\rangle$ is the Bayes posterior on θ based on prior W and data $\vec{Y}\langle 1\rangle, \ldots, \vec{Y}\langle i\rangle$, and $q_W(O_1\langle i\rangle \mid Y_1\langle i\rangle, Y_0\langle i\rangle) := \int q_\theta(O_1\langle i\rangle \mid Y_1\langle i\rangle, Y_0\langle i\rangle) dW(\theta)$ is the Bayes predictive. By multiplying out the conditional probability mass functions r_i, we then get that $M^{\langle n\rangle}_{r,\theta_0} = \prod_{i=1}^{n} M_{r,\theta_0}\langle i\rangle$ is a Bayes factor between the Bayes marginal based on W and θ_0 (GHK explain how such a correspondence between Bayes factors and test martingales

holds more generally for simple null hypotheses). We do not know of a prior for which this Bayes factor or the constituent products have an analytic expression, but it can certainly be implemented using e.g. Gibbs sampling.

As explained underneath (2.12), the use of the r_i instead of q_{θ_1} does not compromise on safety: type-I errors and confidence sequences based on M_{r,θ_0} remain valid, whether the r_i are plug-in estimators or Bayes predictive distributions, no matter what prior W was chosen. Thus, our set-up is actually more intimately related to the concept of *luckiness* in the machine learning theory literature (Grünwald and Mehta, 2019) than to 'pure' Bayesian statistics. The type-I error guarantee always holds, also when the prior is 'misspecified', putting most of its mass in a region of the parameter space far from the actual θ from which the data were sampled. Given a minimal clinically relevant effect size $\theta_{\min}$, the worst case logarithmic growth rate of M_{r,θ_0} will in general be less than that of the GROW $M_{\theta_{\min},\theta_0}$. Nevertheless, M_{r,θ_0} can come quite close to the optimal for a whole range of potentially data-generating θ and may thus sometimes be preferable over choosing $M_{\theta_{\min},\theta_0}$. More precisely, the use of a prior allows us to exploit favourable situations in which θ is even smaller (more extreme) than $\theta_{\min}$. In such situations, the GROW $M_{\theta_{\min},\theta_0}$ is effectively misspecified. By using r_i that learn from the data, we may actually get an E-variable that grows faster than the GROW $M_{\theta_{\min},\theta_0}$ which is fully committed to detecting the worst case $\theta_{\min}$.

In Figure 2.7 we illustrate such a situation where we start with 1000 participants in both groups. We generated data using different hazard ratios, and used a 'misspecified' $M_{\theta_1,1}$ that always used $\theta_1 = 0.8$. Note that while this is still the GROW (minimax optimal) martingale test under $H_1 = \{P_\theta : \theta_1 \leq 0.8\}$, if we knew the true θ, we could use the faster-growing test martingale $M_{\theta,1}$. We will call the test based on this latter martingale the *oracle* exact safe logrank test, since it is based on inaccessible (oracle) knowledge. We estimated the number of events that allows for 80% power for the tests based on $M_{0.8,1}$ and the oracle $M_{\theta,1}$ and the prequential plug-in $M_{r,1}$ with r as in (2.23). In all cases we used the aggressive stopping rule that stops as soon as $M_{\cdot,1} > 1/\alpha = 20$. We see that, as the true θ gets smaller than 0.8, we need less events using the GROW test $M_{0.8,1}$ (the data are favorable to us), but using the oracle exact safe logrank test we get a considerable additional reduction. The prequential plug-in $M_{r,1}$ 'tracks' the oracle $M_{\theta,1}$ by learning the true θ from the data: for θ near 0.8, it behaves worse (more data are needed) than $M_{0.8,1}$ (which knows the right θ from the start), but for $\theta < 0.6$ it starts to behave better. For comparison we also added the methods of Figure 2.6. Notably, the O'Brien-Fleming procedure, even though unsuitable for optional continuation, needs even more events than the misspecified safe logrank test $M_{0.8,1}$ as soon as θ goes below 0.8 (the simulations were performed using exactly the same algorithms as for Figure 2.6 so the y-axis at $\theta = 0.8$ coincides with that of Figure 2.6, but now with absolute rather than relative numbers); details are described in Appendix Section 2.D).

2.4.2 Anytime-valid confidence sequences

Standard tests give rise to confidence intervals by varying the null and 'inverting' the corresponding tests. In analogous fashion, test martingales can be used to derive *anytime-*

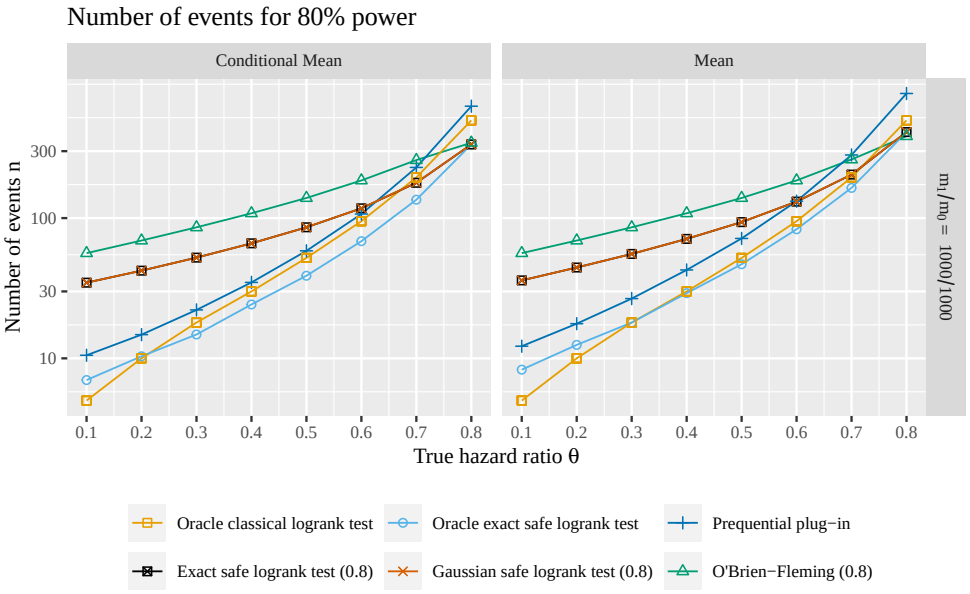

Figure 2.7. *We show the number of events at which one can stop retaining 80% power at* $\alpha = 0.05$ *using the process* M_{θ_1,θ_0} *with* $\theta_0 = 1$ *and* $\theta_1 = 0.80$ *when the true hazard ratio* θ *generating the data is different from* θ_1*. 'Oracle' means that the method is specified with knowledge of the true* θ*, which in reality is unknown. Note that the y-axis is logarithmic.*

valid (AV) confidence sequences (Darling and Robbins, 1967; Lai, 1976; Howard et al., 2018, 2021). In our setting, a $(1-\alpha)$-AV confidence sequence is a sequence of confidence intervals $\{\text{CI}_i\}_{i\in\mathbb{N}}$, one for each consecutive event, such that

$$P_\theta\,(\text{there is an } i \in \mathbb{N} \text{ with } \theta \notin \text{CI}_i) \leq \alpha. \tag{2.25}$$

A standard way to design $(1-\alpha)$- AV confidence sequences, translated to our logrank setting, is to use a prequential plug-in or Bayesian-based r as described in the previous subsection. After observing n events, one reports $\text{CI}_{n,\alpha} = [\theta_{n,L}, \theta_{n,U}]$ where $\theta_{i,L}$ is the largest θ_0 such that for all $\theta' \leq \theta_0$, $M^{\langle n\rangle}_{r,\theta'} \geq 1/\alpha$; similarly $\theta_{n,U}$ is the smallest θ_0 such that for all $\theta' \geq \theta_0$, $M^{\langle n\rangle}_{r,\theta'} \geq 1/\alpha$. That is, we check (2.12) where we vary θ_0 and we report the smallest interval such that $M_{r,\theta_0} > 1/\alpha$ outside this interval. Ville's inequality immediately shows that this gives an AV confidence sequence for arbitrary instantiations of r.

2.4.3 Covariates: the full Cox Proportional Hazards E-Variable

We extend the process of Section 2.1 (for now without ties) to explicitly represent participants, as done above Example 2 (and with the same notation as used there). Additionally, we now also fix a set of d covariates and let $\mathbf{Z}\langle i\rangle = (\vec{z}_1 \ldots \vec{z}_m)\langle i\rangle$ be the matrix consisting of the covariate vectors for each participant at the time of the i^{th} event:

$\vec{z}_j\langle i\rangle = (z_{j,1}\langle i\rangle, \dots, z_{j,d}\langle i\rangle)$. We let random variable $J\langle i\rangle$ denote the index of the patient to which the i^{th} event happens, and consider the extended process $J\langle 1\rangle, J\langle 2\rangle, \dots$ where the information that is available at time i is $\vec{g}, J\langle 1\rangle, \dots, J\langle i\rangle, \mathbf{Z}\langle 1\rangle, \dots, \mathbf{Z}\langle i\rangle$. In this section we re-define $\vec{Y}\langle i\rangle = (Y_1\langle i\rangle, \dots, Y_m\langle i\rangle) \in \{0,1\}^m$ to be a vector of m components, the j component indicating whether the j^{th} participant is still without an event just before the i^{th} event. Accordingly we set $\vec{Y}\langle 1\rangle = (1, \dots, 1)$ and we set $\vec{Y}_j\langle i+1\rangle = \vec{Y}_j\langle i\rangle$ for $j \neq J\langle i\rangle$, $\vec{Y}_j\langle i+1\rangle = 0$ for $j = J\langle i\rangle$. The conditional distribution underlying the process is now denoted $P_{\beta,\theta}$ with $\theta > 0$ and $\beta \in \mathbb{R}^d$, defined as follows: $P_{\beta,\theta}$ is given by, for $j \in \{1, \dots, m\}$ and $\vec{y} \in \{0,1\}^m$ with $\vec{y}_j = 1$:

$$P_{\beta,\theta}(J\langle i\rangle = j \mid \vec{Y}\langle 1\rangle, \mathbf{Z}\langle 1\rangle, \dots, \vec{Y}\langle i\rangle, \mathbf{Z}\langle i\rangle) := P_{\beta,\theta}(J\langle i\rangle = j \mid \vec{Y}\langle i\rangle, \mathbf{Z}\langle i\rangle)$$
$$P_{\beta,\theta}(J\langle i\rangle = j \mid \vec{Y}\langle i\rangle = \vec{y}, \mathbf{Z}\langle i\rangle = \mathbf{z}) := q_{\beta,\theta}(j \mid \vec{y}, \mathbf{z}) := \frac{\exp(\beta^T \vec{z}_j + \theta' g_j)}{\sum_{j':\vec{y}_{j'}=1} \exp(\beta^T \vec{z}_{j'} + \theta' g_{j'})}, \tag{2.26}$$

with $\theta' = \log\theta$ and $\mathbf{z} = (\vec{z}_1, \dots, \vec{z}_m)$. This is consistent with Cox' (1972) proportional hazards regression model: the probability that the j^{th} participant has an event, assuming he/she is still at risk, is proportional to the exponentiated weighted covariates, with group membership being one of the covariates. In case $\beta = 0$, this is easily seen to coincide with the definition of P_θ via (2.6).

E-Variables and Martingales Let W be a prior distribution on $\beta \in \mathbb{R}^d$ for some $d > 0$. (W may be degenerate, i.e. put mass one in a specific parameter vector β_1). We let

$$q_{W,\theta}(j \mid \vec{y}, \mathbf{z}) = \int q_{\beta,\theta}(j \mid \vec{y}, \mathbf{z}) dW(\beta).$$

Consider a measure ρ on $\mathbb{R}^d$ (e.g. Lebesgue or some counting measure) and we let $\mathscr{W}$ be the set of all distributions on $\mathbb{R}^d$ which have a density relative to ρ, and $\mathscr{W}^\circ \subset \mathscr{W}$ be any convex subset of $\mathscr{W}$ (we may take $\mathscr{W}^\circ = \mathscr{W}$, for example). We define $\tilde{q}_{\leftarrow W,\theta_0}(\cdot \mid \vec{y}, \mathbf{z})$ to be the *reverse information projection* (Li, 1999) (RIPr) of $q_{W,\theta}(j \mid \vec{y}, \mathbf{z})$ on $\{q_{W,\theta_0} : W \in \mathscr{W}^\circ\}$ such that

$$D(q_{W,\theta_1}(\cdot \mid \vec{y}, \mathbf{z}) \| \tilde{q}_{\leftarrow W,\theta_0}(\cdot \mid \vec{y}, \mathbf{z})) = \inf_{W' \in \mathscr{W}^\circ} D(q_{W,\theta_1}(\cdot \mid \vec{y}, \mathbf{z}) \| q_{W',\theta_0}(\cdot \mid \vec{y}, \mathbf{z})).$$

We know from Li (1999) and GHK that $\tilde{q}_{\leftarrow W,0}(\cdot \mid \vec{y})$ exists. As explained by GHK in the context of E-variables for 2×2 contingency tables, the fact that the random variables $Y\langle i\rangle$ constituting our random process have finite range implies that, for each W, the infimum is in fact achieved by some distribution W' with finite support on $\mathbb{R}^d$. For given $\theta_0, \theta_1 > 0$, let

$$M_{W,\theta_1,\theta_0}\langle i\rangle = \frac{q_{W,\theta_1}(J\langle i\rangle \mid \vec{Y}\langle i\rangle, \mathbf{Z}\langle i\rangle)}{q_{\leftarrow W,\theta_0}(J\langle i\rangle \mid \vec{Y}\langle i\rangle, \mathbf{Z}\langle i\rangle)} \tag{2.27}$$

be our analogue of $M_{\theta_1,\theta_0}\langle i\rangle$ as in (2.8).

Theorem 2.4.1. [Corollary of Theorem 1 from GHK19] *For every prior W on $\mathbb{R}^d$, for all $\tilde{\beta} \in \mathbb{R}^d$,*

$$\mathbf{E}_{P_{\tilde{\beta},\theta_0}}[M_{W,\theta_1,\theta_0}\langle i\rangle \mid \vec{Y}\langle 1\rangle, \mathbf{Z}\langle 1\rangle, \ldots, \vec{Y}\langle i\rangle, \mathbf{Z}\langle i\rangle] = \sum_{j\in[m]} q_{\beta,\theta_0}(j \mid \vec{Y}\langle i\rangle, \mathbf{Z}\langle i\rangle) \cdot \frac{q_{W,\theta_1}(j \mid \vec{Y}\langle i\rangle, \mathbf{Z}\langle i\rangle)}{q_{\leftarrow W,\theta_0}(j \mid \vec{Y}\langle i\rangle, \mathbf{Z}\langle i\rangle)} \leq 1 \tag{2.28}$$

so that $M_{W,\theta_1,\theta_0}\langle i\rangle$ is an E-variable conditional on $\vec{Y}\langle 1\rangle, \mathbf{Z}\langle 1\rangle, \ldots, \vec{Y}\langle i\rangle, \mathbf{Z}\langle i\rangle$.

Note that the result does not require the prior W to be well-specified in any-way: under any $(\tilde{\beta}, \theta_0)$ in the null distribution, even if $\tilde{\beta}$ is completely disconnected to W, $M_{W,\theta_1,\theta_0}\langle i\rangle$ is an E-variable conditional on past data.

In particular, since the result holds for arbitrary priors, it holds, at the n^{th} event time, for the Bayesian posterior $W_{n+1} = W_1 \mid \vec{Y}\langle 1\rangle, \mathbf{Z}\langle 1\rangle, \ldots, \vec{Y}\langle n\rangle, \mathbf{Z}\langle n\rangle$, based on arbitrary prior W_1 with density w_1, i.e. the density of W_{n+1} is given by

$$w_{n+1}(\beta) \propto \prod_{i=1}^{n} q_{\beta,\theta}(J\langle i\rangle \mid \vec{Y}\langle i\rangle, \mathbf{Z}\langle i\rangle) w_1(\beta).$$

Using Definition Theorem 2.0.3, we can therefore, for each prior W_1, construct a test martingale $M^{\langle n\rangle} := \prod_{i=1}^{n} M_{W_i,\theta_1,\theta_0}$ that 'learns' β from the data, analogously to (2.24), and computes a new RIPr at each event time i.

How to find the RIPr While in general, it is not clear how to calculate the RIPr $q_{\leftarrow W,\theta_0}$, Li (1999); Li and Barron (2000) have designed an efficient algorithm for approximating it, which is feasible as long as we restrict $\mathscr{W}^\circ$ to be the set of all priors W for which, for all $j \in [n]$, $Q_{W,\theta_0}(J\langle i\rangle = j \mid \vec{Y}_j\langle i-1\rangle = 1) \geq \delta$, for $\ell = 1, \ldots, k$, for some $\delta > 0$. The algorithm achieves an approximation error of $O((\log(1/\delta))/M)$ if run for M steps, where each step takes time linear in d. Since the factor is logarithmic in $1/\delta$, we can take a very small value of δ and then the requirement does not seem overly restrictive. Exploring whether the Li-Barron algorithm really allows us to compute the RIPr for the Cox model, and hence M_{W,θ_1,θ_0} in practice, is a major goal for future work.

Ties While in the case without covariates, our E-variables allowing for ties (Section 2.1.2) correspond to a likelihood ratio of noncentral hypergeometrics, the situation is not so simple if there are covariates—although deriving the appropriate extension of the non-central hypergeometric partial likelihood is possible, one ends up with a hard-to-calculate formula (Peto, 1972). Various approximations have been proposed in the literature (Cox, 1972; Efron, 1974). In case these preserve the E-variable and martingale properties, they would retain type-I error probabilities under optional stopping and we could use them without problems. We do not know whether this is the case however; for the time being,

we recommend handling ties by putting the events in a worst-case order, leading to the smallest values of the E-variable of interest, as this is bound to preserve the type-I error guarantees.

2.5 Discussion, Conclusion and Future Work

We introduced the safe logrank test, a version of the logrank test that can retain type-I error guarantees under optional stopping and continuation. Extensive simulations revealed that, if we do engage in optional stopping, it is competitive with the classical logrank test (which neither allows in-trial optional stopping nor optional continuation) and α-spending (which allows forms of optional stopping but not optional continuation). We provided an approximate test for applications in which only summary statistics are available and also showed how the safe logrank test can be used in combination with (informative) prior and prequential learning approaches, when no effect size of minimal clinical relevance can be specified. Two of our extensions invite further research: We introduced anytime-valid confidence sequences for the hazard ratio, and will study these more in future work into their performance in comparison to other approaches. We also introduced an extension to Cox' proportional hazards regression, which promises type-I error guarantees even if the alternative model is equipped with arbitrary priors. In future work, we plan to implement this extension – which requires the use of sophisticated methods for estimating mixture models. The GROW safe logrank tests (exact and Gaussian) are already available in our SafeStats R package (Turner et al., 2022). We end with two final points of discussion: *staggered entries* and *doomed trials*.

Staggered entry Earlier approaches to sequential time-to-event analysis were also studied under scenarios of staggered entry, where each patient has its own event time (e.g. time to death since surgery), but patients do not enter the follow-up simultaneously (such that the risk set of e.g. a two-day-after-surgery event changes when new participants enter and survive two days). Sellke and Siegmund (1983) and Slud (1984) show that, in general, martingale properties cannot be preserved under such staggered entry, but that asymptotic results are hopeful (Sellke and Siegmund, 1983) as long as certain scenarios are excluded (Slud, 1984). When all participants' risk is on the same (calendar) time scale (e.g. infection risk in a pandemic; staggered entry now amounts to left-truncation, which we can deal with), or new patients enter in large groups (allowing us to stratify), staggered entry poses no problem for our methods. But research is still ongoing into those scenarios in which our inference is fully safe for patient time under staggered entry, and those that need extra care.

Your trial is not doomed In their summary of conditional power approaches in sequential analysis Proschan et al. (2006) write that low conditional power makes a trial futile. Continuing a trial in such case could only be worth the effort to rule out an effect of clinical relevance, when the effect can be estimated with enough precision. However, if "both conditional and revised unconditional power are low, the trial is doomed because a null result is both likely and uninformative" (Proschan et al., 2006, p. 63). While this is the

case for all existing sequential approaches that set a maximum sample size, this is not the case for safe tests. Any trial can be extended and possibly achieve 100% power or in an anytime-valid confidence sequence show that the effect is too small to be of interest. This is especially useful for time-to-event data when sample size can increase by extending the follow-up of the trial, without recruiting more participants. Moreover, new participants can always be enrolled either within the same trial or by spurring new trials that can be combined indefinitely in a cumulative meta-analysis.

Code availability

This chapter's R code is available on `https://osf.io/3n8g2/` (Ter Schure and Pérez-Ortiz, 2022).

Appendices

Appendix Section 2.A connects the simple risk set processes in this chapter to a continuous time survival process. Section 2.B gives an additional argument for the GROW criterion, Section 2.C gives a more detailed derivation of the logrank test as a score test for single events and ties and Section 2.D gives a step-by-step account of the sample size comparison simulations.

2.A Towards Continuous Time

In Section 2.1.1 and Section 2.4.3, the expression (2.26) and its simplified form (2.7) defining the safe logrank test appeared as a factor in a standard likelihood, defined relative to a basic stochastic process in which the time between events was not formalized. Cox (1975) simply claimed it to be also a partial likelihood for the underlying continuous-time process with proportional hazards determined by (β, θ'). Slud (1992) proved that it can indeed be seen as such, in the same general setting with time-varying continuous covariates that we consider here (see also Andersen et al. (1993) for closely related results). This already shows that the test martingales of Section 2.1.1 (no covariates, no ties) and Section 2.4.3 (covariates, no ties) also remain test martingales in the continuous time setting. The only part that is not covered by such existing results is the case of ties, Section 2.1.2: if we plug the conditional distributions (2.15) that allow for ties into the test martingale (2.8), do we still get a test martingale in continuous time if $\theta = 1$? Since we have not been able to find a complete proof in the literature of this result or a result that would directly imply it, we derive a simple version of such a result below, implicitly also proving the result without ties, though in a less general setting (without covariates, and assuming continuous hazards) than Slud. For simplicity we only treat the case without censoring again.

As in Section 2.4.3 and above Example 2, we identify each of the $m = m_0 + m_1$ participants by an index $j \in [m] := \{1, \dots, m\}$, with $\vec{g} \in \{0,1\}^m$, and $\vec{g}_j \in \{0,1\}$ denoting the group assignment of participant j. However, for simplicity we shall not make use of any covariates. For $t \in \mathbb{R}_0^+$ (continuous time), we let $T_j = t$ denote that the j^{th} participant had an event at time j; $T\langle i \rangle$ denotes the time at which the i-th event takes place. For $g \in \{0,1\}$, we let $Y_g(t) = \sum_{j:\vec{g}_j = g} Y_j(t)$ be the number of participants at risk in the group g at time t.

Fix some time $t^* > 0$. If patient with index j is in group g, then by assumption (2.13) we have, for fixed $0 < \epsilon < t^*$,

$$p_g^*(\epsilon) := P_g(t^* - \epsilon < T_j \leq t^* \mid T_j > t^* - \epsilon) = 1 - \exp\left(\int_{t^*-\epsilon}^{t^*} \lambda_g(s)\,\mathrm{d}s\right).$$

Let O_g^* be the number of patients of group $g \in \{0,1\}$ that witnessed the event of interest in the time interval $(t^* - \epsilon, t^*]$, and let $O^* = O_0^* + O_1^*$ be the total number of such patients.

Let $\mathscr{E}$ be the event denoting everything that is known just before time $t^*-\epsilon$, i.e. if k events happened before $t^*-\epsilon$, then $\mathscr{E}$ is the event that

$$T\langle 1\rangle = t_1, \ldots, T\langle k\rangle = t_k, Y_0\langle 1\rangle = y_0\langle 1\rangle, Y_1\langle 1\rangle = y_1\langle 1\rangle, \ldots, Y_0\langle k\rangle = y_0\langle k\rangle, Y_1\langle k\rangle = y_1\langle k\rangle$$

for specific $t_1, \ldots, t_k, y_0\langle 1\rangle, y_1\langle 1\rangle, \ldots, y_0\langle k\rangle, y_1\langle k\rangle$. Let $y_g := Y_g\langle k+1\rangle$ and note that y_g can be calculated at all times after $T\langle k\rangle$. By independence of the T_j, the distribution of O_g^* given $\mathscr{E}$ is binomial, given by, for ϵ, t^* such that $t^*-\epsilon > t_k$ and $o_g \leq y_g$,

$$P(O_0^* = o_0, O_1^* = o_1 \mid \mathscr{E}) = \\ \binom{y_0}{o_0} \cdot p_0^*(\epsilon)^{o_0}(1-p_0^*(\epsilon))^{y_0-o_0} \cdot \binom{y_1}{o_1} \cdot p_1^*(\epsilon)^{o_1}(1-p_1^*(\epsilon))^{y_1-o_1}.$$

Now let $\omega_g^*(\epsilon) = p_g^*(\epsilon)/(1-p_g^*(\epsilon))$. As is well known, the probability of (O_0^*, O_1^*) given $\mathscr{E}$ being binomial as above, the conditional probability of observing a particular O_1^* given $O^* = O_0^* + O_1^*$ and $\mathscr{E}$ must be given by Fisher's non-central hypergeometric distribution with parameter $\omega^*(\epsilon) = \omega_1^*(\epsilon)/\omega_0^*(\epsilon)$, whose probability mass function is given by, with ω abbreviating $\omega^*(\epsilon)$,

$$P(O_1^* = o_1 \mid \mathscr{E}, O^* = o) = q_\omega(o_1 | y_0, y_1, o) = \frac{\binom{y_1}{o_1}\binom{y_0}{o-o_1}\omega^{o_1}}{\sum_{\max\{0,o-y_1\}\leq u\leq \min\{y_1,o\}} \binom{y_1}{u}\binom{y_0}{o-u}\omega^u}.$$

We now note that if $\theta = 1$, then $\omega^*(\epsilon) = 1$ irrespective of ϵ and this reduces to the hypergeometric distribution with parameter $\theta = 1$ as in (2.15). This proves the claim made in Section 2.1.2: q_θ as in (2.15) with $\theta = 1$ is the correct conditional distribution under the continuous time model introduced above Example 2 with hazard ratio 1. It follows, using the reasoning in Example 1, that $M_{\theta_1,1}^{\langle I\rangle}$ as underneath (2.15) is a test martingale, and our test with $\theta_0 = 1$ is once again exact. Under hazard ratios $\theta \neq 1$, $\omega^*(\epsilon) \neq \theta$, so we do not have an exact supermartingale (and hence not an exact test) anymore. Still, as noted by Mehrotra and Roth (2001),

$$\lim_{\epsilon \downarrow 0} \omega^*(\epsilon) = \theta, \tag{2.A.1}$$

so, if we make observations at subsequent time points that are close enough to each other, all of our results still hold in an approximate sense.

From (2.A.1) we also see that

$$P(O_1\langle k+1\rangle = 1 \mid \mathscr{E}, T\langle k+1\rangle = t^*) = \lim_{\epsilon \downarrow 0} P(O_1^* = 1 \mid \mathscr{E}, O^* = 1) = q_\theta(o_1 \mid y_0, y_1),$$

with q_θ given as in (2.6). This shows that our original conditional distributions are correct as well in continuous time, now under each hazard ratio $\theta > 0$, not just $\theta = 1$, and $M_{r,\theta}$ gives an exact test martingale and hence an exact test for each θ, as long as there are no ties.

2.B Expected Stopping Time, GROW and Wald's Identity

Let P_{θ_0} represent our null model, and let, as before, the alternative model be given as $H_1 = \{P_{\theta_1} : \theta_1 \in \Theta_1\}$ with $\Theta_1 = \{\theta_1 : 0 < \theta_1 \leq \theta_{\min}\}$ for some $\theta_{\min} < 1$. Suppose we perform a level α test based on a test martingale M_{θ_1,θ_0} using the aggressive stopping rule: stop as soon as $M_{\theta_1,\theta_0} \geq 1/\alpha$. The GROW criterion (Chapter 2, Example 2) tells us to use $\theta_1 = \theta_{\min}$. Here we motivate this GROW criterion by showing that it minimizes, in a worst-case sense, the expected number of events needed before there is sufficient evidence to stop. The calculation below ignores the practical need to prepare for a bounded maximum number of events. For such more complicated considerations, we need to resort to simulations as in the main text.

We will further make the simplifying assumption that the initial risk sets (i.e. m_0 and m_1) are large enough so that for all sample sizes we will ever encounter, $Y_0\langle i\rangle / Y_1\langle i\rangle \approx m_0/m_1$. This allow us to act as if the random variables $O_1\langle i\rangle$ are i.i.d. Bernoulli: the ith event is sampled from $q_{\theta_1}\big(O_1\langle i\rangle \mid (Y_0\langle i\rangle, Y_1\langle i\rangle)\big)$ for some $\theta_1 \in \Theta_1$, with q_{θ_1} as given by (2.6), which becomes independent of i if y_0/y_1 is replaced by m_0/m_1. Assuming i.i.d. data enables a standard argument based on Wald's (1947) identity, originally due to Breiman (1961). As said, we stop as soon as $M := M_{\theta_1,\theta_0} \geq 1/\alpha$ or when we run out of data, leading to a stopping time τ_{θ_1} (which we will denote by τ when θ_1 is clear from the context). Suppose first that we happen to know that the data comes from a specific $\theta \in \Theta_1$. Wald's identity now gives:

$$\mathbf{E}_{P'_\theta}[\tau] = \frac{\mathbf{E}_{P'_\theta}[\log M^{\langle\tau\rangle}_{\theta_1,\theta_0}]}{\mathbf{E}_{P'_\theta}[\log M_{\theta_1,\theta_0}\langle 1\rangle]}.$$

For simplicity we will further assume that the number of people at risk is large enough so that the probability that we run out of data before we can reject is negligible. The right-hand side can then be further rewritten as

$$\frac{\mathbf{E}_{P'_\theta}[\log M^{\langle\tau\rangle}_{\theta_1,\theta_0}]}{\mathbf{E}_{P'_\theta}\left[\log \frac{p'_{\theta_1}(O_1\langle 1\rangle)}{p'_{\theta_0}(O_1\langle 1\rangle)}\right]} = \frac{\log\frac{1}{\alpha} + \text{VERY SMALL}}{\mathbf{E}_{P'_\theta}\left[\log \frac{p'_{\theta_1}(O_1\langle 1\rangle)}{p'_{\theta_0}(O_1\langle 1\rangle)}\right]} \tag{2.B.1}$$

with VERY SMALL between 0 and $\log|\theta_1/\theta_0|$, and $p'_{\theta_1}(O_1\langle 1\rangle) = q_{\theta_1}(O_1\langle 1\rangle \mid Y_1\langle 1\rangle, Y_0\langle 1\rangle)$. The first equality is just definition, the second follows because we reject *as soon as* $M^{\langle\tau\rangle}_{\theta_1,\theta_0} \geq 1/\alpha$, so $M^{\langle\tau\rangle}_{\theta_1,\theta_0}$ can't be smaller than $1/\alpha$, and it can't be larger by more than a factor equal to the maximum likelihood ratio at a single outcome (if we would not ignore the probability of stopping because we run out of data, there would be an additional small term in the numerator).

If we try to find the θ_1 which minimizes this, and – as is customary in sequential analysis – we approximate the minimum by ignoring the VERY SMALL part, we see that the expression is minimized by maximizing $\mathbf{E}_{P'_\theta}[\log \frac{p_{\theta_1}(Y\langle 1\rangle)}{p_{\theta_0}(Y\langle 1\rangle)}]$ over θ_1. The maximum is clearly achieved by $\theta_1 = \theta$; the expression in the denominator then becomes the KL divergence between two Bernoulli distributions. It follows that under θ, the expected number of outcomes

until rejection is minimized if we set $\theta_1 = \theta$. Thus, we use the GROW E-variable relative to $\{\theta\}$ as our actual E-variable. We still need to consider the case that, since the real H_1 is 'composite', as statisticians, we do not know the actual θ; we only know $0 < \theta \leq \theta_{\min}$. So we might want to take a worst-case approach and use the θ_1 achieving

$$\max_{\theta_1} \min_{\theta:0<\theta\leq\theta_{\min}} \mathbf{E}_{P'_\theta}\left[\log \frac{p'_{\theta_1}(O_1\langle 1\rangle)}{p'_{\theta_0}(O_1\langle 1\rangle)}\right],$$

since, repeating the reasoning leading to (2.B.1), this θ_1 should be close to achieving

$$\min_{\theta_1} \max_{\theta:0<\theta\leq\theta_{\min}} \mathbf{E}_{P'_\theta}[\tau_{\theta_1}]$$

But this just tells us to use the GROW E-variable relative to H_1, which is what we were arguing for.

2.C Logrank test as a score test

2.C.1 Logrank test statistic for single events and ties

Let $Y_1\langle i\rangle$ denote that number of participants in the risk set that are in the treatment group at the time of the i^{th} event, and analogously $Y_0\langle i\rangle$ for the number of participants at risk in the control group. Let $O_1\langle i\rangle$ and $O_0\langle i\rangle$ count the number of observed events in the treatment group and control group at the i^{th} event time. For single events $O_1\langle i\rangle = 1$ if the event occurred in the treatment group, and $O_1\langle i\rangle = 0$ if a single event occurred in the control group. We can extend this Bernoulli case to multiple simultaneous events (ties) in which case $O_1\langle i\rangle$ can be larger than 1. Here we discuss both cases (single event and ties) together, but we will discuss them separately later on. We define $Y\langle i\rangle = Y_1\langle i\rangle + Y_0\langle i\rangle$ and $O\langle i\rangle = O_1\langle i\rangle + O_0\langle i\rangle$.

Under the null hypothesis, at each event time i, the number of observed events in the treatment group $O_1\langle i\rangle$ follows a hypergeometric distribution with an expected number of events $E_1\langle i\rangle$ and a variance $V_1\langle i\rangle$ that depends on the risk set as follows:

$$A_1\langle i\rangle = \frac{Y_1\langle i\rangle}{Y\langle i\rangle}, \qquad E_1\langle i\rangle = O\langle i\rangle \cdot A_1\langle i\rangle, \qquad V_1\langle i\rangle = O\langle i\rangle \cdot A_1\langle i\rangle \cdot \big(1 - A_1\langle i\rangle\big) \cdot \frac{Y\langle i\rangle - O\langle i\rangle}{Y\langle i\rangle - 1}.$$

This is the formulation found in Cox (1972, equation (26)) with $\frac{Y\langle i\rangle - O\langle i\rangle}{Y\langle i\rangle - 1}$ a "multiplicity" correction or "correction for ties" (Klein and Moeschberger, 2006, p. 207). In case of single events with $O\langle i\rangle = 1$ the variance reduces to the Bernoulli variance:

$$V_1\langle i\rangle = A_1\langle i\rangle \cdot \big(1 - A_1\langle i\rangle\big) = E_1\langle i\rangle \cdot \big(1 - E_1\langle i\rangle\big).$$

As a test statistic, a logrank Z-statistic is constructed that is calculated either for the treatment or for the control group and has an approximate standard normal (Gaussian) distribution under the null hypothesis. This Z-statistic for the treatment group 1 is:

$$Z = \frac{\sum_i \{O_1\langle i\rangle - E_1\langle i\rangle\}}{\sqrt{\sum_i V_1\langle i\rangle}}.$$

This test statistic is used to reject the null hypothesis at level α in favor of a one-sided alternative (e.g. $H_0 : \lambda_1(t\langle i\rangle) < \lambda_0(t\langle i\rangle)$ for some i) in case the value of the Z-statistic is smaller than z_α. You expect a negative value for the Z-statistic for a lower risk in the treatment group, so you need the α^{th} lower percentage point of the standard normal (Gaussian) distribution. A two-sided test can be constructed by comparing $|Z| \geq z_{\alpha/2}$.

2.C.2 Score test for the Bernoulli partial likelihood (single events)

In Section 2.1.1 we constructed a partial likelihood based on the probability mass function of a Bernoulli $y_1\theta/(y_0 + y_1\theta)$-distribution. Given an observed data set of participants at risk ($\vec{Y}\langle i\rangle$) at all event times i we can define a product of Bernoulli likelihoods. Following Cox (1972), we define these in terms of the logarithm of the hazard ratio, i.e. $\beta = \log\theta$ with $\theta = \lambda_1(t\langle i\rangle)/\lambda_1(t\langle i\rangle)$ for all event times $t\langle i\rangle$:

$$\mathfrak{L}\big(\beta \mid \vec{Y}\langle 1\rangle, \vec{Y}\langle 2\rangle, \ldots, \vec{Y}\langle n\rangle\big) = \prod_{i=1}^{n} q_\beta\big(O_1\langle i\rangle \mid (Y_0\langle i\rangle, Y_1\langle i\rangle)\big),$$

where

$$q_\beta(O_1\langle i\rangle \mid (Y_0\langle i\rangle, Y_1\langle i\rangle)) = \left(\frac{Y_1\langle i\rangle \cdot \exp(\beta)}{Y_0\langle i\rangle + Y_1\langle i\rangle \cdot \exp(\beta)}\right)^{O_1\langle i\rangle} \left(\frac{Y_0\langle i\rangle}{Y_0\langle i\rangle + Y_1\langle i\rangle \cdot \exp(\beta)}\right)^{1-O_1\langle i\rangle}.$$

This is the likelihood formulated for single events by Cox (1972) for his proportional hazards model in the two-sample case (a single categorical covariate indicating two groups).

Following the likelihood above, our loglikelihood is:

$$\begin{aligned} L(\beta \mid \vec{Y}\langle 1\rangle, \vec{Y}\langle 2\rangle, \ldots, \vec{Y}\langle n\rangle) &= \sum_{i=1}^{n} \Big\{ O_1\langle i\rangle \cdot \Big(\log[Y_1\langle i\rangle] + \beta - \log\big[Y_0\langle i\rangle + Y_1\langle i\rangle \cdot \exp(\beta)\big]\Big) \\ &\qquad + \big(1 - O_1\langle i\rangle\big) \cdot \Big(\log[Y_0\langle i\rangle] - \log[Y_0\langle i\rangle + Y_1\langle i\rangle \cdot \exp(\beta)]\Big)\Big\} \\ &= \sum_{i=1}^{n} \Big\{ O_1\langle i\rangle\beta + O_1\langle i\rangle \cdot \Big(\log[Y_1\langle i\rangle] - \log[Y_0\langle i\rangle]\Big) + \log[Y_0\langle i\rangle] \\ &\qquad - \log\big[Y_0\langle i\rangle + Y_1\langle i\rangle \cdot \exp(\beta)\big]\Big\}. \end{aligned}$$

If we omit the parts that do not depend on β, we get the expression in Cox (1972):

$$L(\beta \mid \vec{Y}\langle 1\rangle, \vec{Y}\langle 2\rangle, \ldots, \vec{Y}\langle n\rangle) \propto \sum_{i=1}^{n} \Big\{ O_1\langle i\rangle\beta - \log\big[Y_0\langle i\rangle + Y_1\langle i\rangle \cdot \exp(\beta)\big]\Big\}.$$

So if we take the score:

$$U(\beta) = \frac{\mathrm{d}L(\beta \mid \vec{Y}\langle 1\rangle, \vec{Y}\langle 2\rangle, \ldots, \vec{Y}\langle n\rangle)}{\mathrm{d}\beta} = \sum_{i=1}^{n} \left\{ O_1\langle i\rangle - \frac{Y_1\langle i\rangle \cdot \exp(\beta)}{Y_0\langle i\rangle + Y_1\langle i\rangle \cdot \exp(\beta)} \right\},$$

with variance (by the observed Fisher information)

$$
\begin{aligned}
-U'(\beta) &= -\frac{\mathrm{d}^2 L(\beta \mid \vec{Y}\langle 1\rangle, \vec{Y}\langle 2\rangle, \ldots, \vec{Y}\langle n\rangle)}{\mathrm{d}\beta^2} \\
&= \sum_{i=1}^{n} \left\{ \frac{\left(Y_0\langle i\rangle + Y_1\langle i\rangle \cdot \exp(\beta)\right) \cdot Y_1\langle i\rangle \cdot \exp(\beta) - Y_1\langle i\rangle \cdot \exp(\beta) \cdot Y_1\langle i\rangle \cdot \exp(\beta)}{\left(Y_0\langle i\rangle + Y_1\langle i\rangle \cdot \exp(\beta)\right)^2} \right\} \\
&= \sum_{i=1}^{n} \left\{ \frac{Y_0\langle i\rangle Y_1\langle i\rangle \exp(\beta)}{(Y_0\langle i\rangle + Y_1\langle i\rangle \exp(\beta))^2} \right\} = \sum_{i=1}^{n} \left\{ \frac{Y_1\langle i\rangle \exp(\beta)}{Y_0\langle i\rangle + Y_1\langle i\rangle \exp(\beta)} \frac{Y_0\langle i\rangle}{Y_0\langle i\rangle + Y_1\langle i\rangle \exp(\beta)} \right\}
\end{aligned}
$$

such that, if we evaluate the score function at a hazard ratio θ of 1, so with $\beta = \exp(\theta) = \exp(1) = 0$, we get:

$$
\begin{aligned}
U(0) &= \sum_{i=1}^{n} \left\{ O_1\langle i\rangle - \frac{Y_1\langle i\rangle}{Y_0\langle i\rangle + Y_1\langle i\rangle} \right\} = \sum_{i=1}^{n} \left\{ O_1\langle i\rangle - \frac{Y_1\langle i\rangle}{Y\langle i\rangle} \right\} = \sum_{i=1}^{n} \{O_1\langle i\rangle - A_1\langle i\rangle\}, \\
-U'(0) &= \sum_{i=1}^{n} \frac{Y_1\langle i\rangle}{Y_0\langle i\rangle + Y_1\langle i\rangle} \cdot \frac{Y_0\langle i\rangle}{Y_0\langle i\rangle + Y_1\langle i\rangle} = \sum_{i=1}^{n} A_1\langle i\rangle \cdot (1 - A_1\langle i\rangle).
\end{aligned}
$$

$A_1\langle i\rangle = E_1\langle i\rangle / O\langle i\rangle = E_1\langle i\rangle$ in case of single events. Hence by standardizing the score function under the null hypothesis, we get the logrank statistic for single events:

$$
Z = \frac{\sum_{i=1}^{n} \{O_1\langle i\rangle - A_1\langle i\rangle\}}{\sqrt{\sum_{i=1}^{n} A_1\langle i\rangle \cdot (1 - A_1\langle i\rangle)}} = \frac{\sum_{i=1}^{n} \{O_1\langle i\rangle - E_1\langle i\rangle\}}{\sqrt{\sum_i V_1\langle i\rangle}}.
$$

2.C.3 Score test for the Fisher hypergeometric partial likelihood (tied events)

The Peto one-step estimator for odds ratios is build from a general score test for 2x2 tables. The appendix of Yusuf et al. (1985) gives description that we can use for the logrank score test in the case of ties.

If at the i^{th} event time $O\langle i\rangle$ events are observed, the expression in (2.15) describes the probability that $O_1\langle i\rangle$ of these events happen in the treatment group. The general description of the likelihood for the i^{th} event time that follows from this is:

$$
\begin{aligned}
s_\beta(O_1\langle i\rangle \mid (Y_0\langle i\rangle, Y_1\langle i\rangle), O\langle i\rangle) &:= \frac{p_0(O_1\langle i\rangle) \cdot \exp(O_1\langle i\rangle \cdot \beta)}{\sum_{k=O_1^{\min}\langle i\rangle}^{O_1^{\max}\langle i\rangle} p_0(k) \cdot \exp(k \cdot \beta)} \\
\text{with } O_1^{\min}\langle i\rangle = \max\{0, O\langle i\rangle - Y_0\langle i\rangle\}; \quad O_1^{\max}\langle i\rangle &= \min\{O\langle i\rangle, Y_1\langle i\rangle\},
\end{aligned}
$$

with $p_0(O_1\langle i\rangle)$ the null hypothesis probability of $O_1\langle i\rangle$ observed events in the treatment group. Our null hypothesis is a hypergeometric distribution:

$$
p_0(O_1\langle i\rangle) = P_0\left((O_1\langle i\rangle \mid (Y_0\langle i\rangle, Y_1\langle i\rangle), O\langle i\rangle\right) = \frac{\binom{Y_1\langle i\rangle}{O_1\langle i\rangle} \cdot \binom{Y_0\langle i\rangle}{O\langle i\rangle - O_1\langle i\rangle}}{\binom{Y\langle i\rangle}{O\langle i\rangle}},
$$

such that $s_\beta = q_\beta$, our Fisher hypergeometric likelihood (2.15) in Section 2.1.2:

$$\begin{aligned} s_\beta(O_1\langle i\rangle \mid (Y_0\langle i\rangle, Y_1\langle i\rangle), O\langle i\rangle) &= \frac{p_0(O_1\langle i\rangle)\cdot \exp(O_1\langle i\rangle\cdot\beta)}{\sum_{k=O_1^{\min}\langle i\rangle}^{O_1^{\max}\langle i\rangle} p_0(k)\cdot\exp(k\cdot\beta)} \\ &= \frac{\frac{\binom{Y_1\langle i\rangle}{O_1\langle i\rangle}\cdot\binom{Y_0\langle i\rangle}{O\langle i\rangle - O_1\langle i\rangle}}{\binom{Y\langle i\rangle}{O\langle i\rangle}}\cdot\exp(O_1\langle i\rangle\cdot\beta)}{\sum_{k=O_1^{\min}\langle i\rangle}^{O_1^{\max}\langle i\rangle}\frac{\binom{Y_1\langle i\rangle}{k}\cdot\binom{Y_0\langle i\rangle}{O\langle i\rangle - k}}{\binom{Y\langle i\rangle}{O\langle i\rangle}}\cdot\exp(k\cdot\beta)} \\ &= \frac{\binom{Y_1\langle i\rangle}{O_1\langle i\rangle}\cdot\binom{Y_0\langle i\rangle}{O\langle i\rangle - O_1\langle i\rangle}\cdot\exp(O_1\langle i\rangle\cdot\beta)}{\sum_{k=O_1^{\min}\langle i\rangle}^{O_1^{\max}\langle i\rangle}\binom{Y_1\langle i\rangle}{k}\cdot\binom{Y_0\langle i\rangle}{O\langle i\rangle - k}\cdot\exp(k\cdot\beta)} \\ &= q_\beta(O_1\langle i\rangle \mid (Y_0\langle i\rangle, Y_1\langle i\rangle), O\langle i\rangle). \end{aligned}$$

So our likelihood is:

$$\mathfrak{L}\left(\beta \mid \vec{Y}\langle 1\rangle, \vec{Y}\langle 2\rangle, \ldots\right) = \prod_i s_\beta(O_1\langle i\rangle \mid (Y_0\langle i\rangle, Y_1\langle i\rangle), O\langle i\rangle)$$

And the loglikelihood is:

$$L(\beta \mid \vec{Y}\langle 1\rangle, \vec{Y}\langle 2\rangle, \ldots) = \sum_i \left\{ \log\left[p_0(O_1\langle i\rangle)\right] + O_1\langle i\rangle\cdot\beta - \log\left[\sum_{k=O_1^{\min}\langle i\rangle}^{O_1^{\max}\langle i\rangle} p_0(k)\cdot\exp(k\cdot\beta)\right]\right\}.$$

And its score:

$$U(\beta) = \frac{\mathrm{d}L(\beta \mid \vec{Y}\langle 1\rangle, \vec{Y}\langle 2\rangle, \ldots)}{\mathrm{d}\beta} = \sum_i \left\{ O_1\langle i\rangle - \frac{\sum_{k=O_1^{\min}\langle i\rangle}^{O_1^{\max}\langle i\rangle} p_0(k)\cdot k\cdot\exp(k\cdot\beta)}{\sum_{k=O_1^{\min}\langle i\rangle}^{O_1^{\max}\langle i\rangle} p_0(k)\cdot\exp(k\cdot\beta)}\right\},$$

with variance (by the observed Fisher information)

$$\begin{aligned} -U'(\beta) &= -\frac{\mathrm{d}^2 L(\beta \mid \vec{Y}\langle 1\rangle, \vec{Y}\langle 2\rangle, \ldots)}{\mathrm{d}\beta^2} \\ &= \sum_i \left\{ \frac{\sum_k p_0(k)\cdot\exp(k\cdot\beta)\cdot\sum_k p_0(k)\cdot k^2\cdot\exp(k\cdot\beta) - \left(\sum_k p_0(k)\cdot k\cdot\exp(k\cdot\beta)\right)^2}{\left(\sum_{k=O_1^{\min}\langle i\rangle}^{O_1^{\max}\langle i\rangle} p_0(k)\cdot\exp(k\cdot\beta)\right)^2}\right\}. \end{aligned}$$

If we evaluate the score function at a hazard ratio θ of 1, so with $\beta = \exp(\theta) = \exp(1) = 0$,

we get:

$$
\begin{aligned}
U(0) &= \sum_i \left\{ O_1\langle i\rangle - \frac{\sum_{k=O_1^{\min}\langle i\rangle}^{O_1^{\max}\langle i\rangle} p_0(k)\cdot k}{\sum_{k=O_1^{\min}\langle i\rangle}^{O_1^{\max}\langle i\rangle} p_0(k)} \right\} = \sum_i \left\{ O_1\langle i\rangle - \sum_{k=O_1^{\min}\langle i\rangle}^{O_1^{\max}\langle i\rangle} p_0(k)\cdot k \right\} \\
&= \sum_i \left\{ O_1\langle i\rangle - O_1\langle i\rangle \cdot \frac{Y_1\langle i\rangle}{Y\langle i\rangle} \right\} = \sum_i \{O_1\langle i\rangle - E_1\langle i\rangle\} \\
U'(O) &= \sum_i \left\{ \frac{\sum_k p_0(k)\cdot\sum_k p_0(k)\cdot k^2 - \left(\sum_k p_0(k)\cdot k\right)^2}{\left(\sum_{k=O_1^{\min}\langle i\rangle}^{O_1^{\max}\langle i\rangle} p_0(k)\right)^2} \right\} \\
&= \sum_i \left\{ \sum_k p_0(k)\cdot k^2 - \left(\sum_k p_0(k)\cdot k\right)^2 \right\} = \sum_i V_1\langle i\rangle.
\end{aligned}
$$

2.D Details of sample size comparison simulations

In this section we lay out the procedure that we used to estimate the expected and maximum number of events required to achieve a predefined power as shown in Figure 2.6 and Figure 2.7 in Section 2.3 and Section 2.4. First we describe how we sampled the survival processes under a specific hazard ratio. We then describe how we estimated the maximum and expected sample size required to achieve a predefined power (80% in our case) for any of the test martingales that we considered (that of the exact safe logrank, its Gaussian approximation, and its prequential-plugin variation). Finally, we explain how the numbers for the classical logrank test and the O'Brien-Fleming procedure were obtained.

In order to simulate the order in which the events in a survival processes happens, we used the risk set process from Section 2.1.1. Indeed, if we are testing some fixed θ_1 with $\theta_1 \leq 1$ against $\theta_0 = 1$, the odds of next event at the i^{th} event time happening in group 1 are $\theta_1 Y_1\langle i\rangle : Y_0\langle i\rangle$ under the alternative hypothesis, which is the hypothesis we sample from. Thus, simulating in which group the next event happens only takes a (biased) coin flip. We consider the one-sided testing scenario θ_1 (for some $\theta_1 \in (0,1)$) vs. $\theta_0 = 1$, and we fix our desired level to $\alpha = 0.05$. For each test martingale $M_{\theta_1,1}^{\langle n\rangle}$ of interest we first consider the stopping rule $\tau = \inf\{n : M_{\theta_1,1}^{\langle n\rangle} \geq 1/\alpha\}$, that is, we stop as soon as $M_{\theta_1,1}^{\langle n\rangle}$ crosses the threshold $1/\alpha$.

In order to estimate the maximum number of events needed to achieve a predefined power with a given test martingale, we turned our attention to a modified stopping rule τ'. Under τ' we stop at the first of two moments: either when our test martingale $M_{\theta_1,1}^{\langle n\rangle}$ crosses the threshold $1/\alpha$ (i.e. at τ) or once we have witnessed a predefined maximum number of events $n_{\max}$. More compactly, this means using the stopping rule τ' given by $\tau' = \min(\tau, n_{\max})$. In those cases in which the test based on the stopping rule τ achieves a power higher than $1-\beta$ (for a type-II error rate β), a maximum number of events

$n_{\max}$ smaller than the initial size of the combined risk groups can be selected to achieve approximate power $1-\beta$ using the rule τ'. A quick computation shows that $n_{\max}$ has the following property: it is the smallest number of events n such that stopping after n events has probability smaller than $1-\beta$ under the alternative hypothesis, that is,

$$P_{\theta_1}(\tau \geq n) \leq 1-\beta.$$

More succinctly, $n_{\max}$ is the (approximate) $(1-\beta)$-quantile of the stopping time τ and can in consequence be estimated experimentally in a straightforward manner.

In order to estimate $n_{\max}$ for a given risk-set sizes m_1, m_0 and alternative hypothesis hazard ratio θ_1, we sampled 10^4 realizations of the survival process (under θ_1) using the method described at the beginning of this section. This allowed us to obtain the same number of realizations of the stopping time τ. We then computed the $(1-\beta)$-quantile of the observed empirical distribution of τ, and reported it as an estimate of the number of events $n_{\max}$ in the 'maximum' column in Figure 2.6.

We assessed the uncertainty in the estimation $n_{\max}$ using the bootstrap. We performed 1000 bootstrap rounds on the sampled empirical distribution of τ, and found that the number of realizations that we sampled (10^4) was high enough so that plotting the uncertainty estimates was not meaningful relative to the scale of our plots. For this reason we omitted the error bars in Figure 2.6 and Figure 2.7.

In the 'mean' column of Figure 2.6 and Figure 2.7 we plotted an estimate of the expected number of events $\tau' = \min(\tau, n_{\max})$. For this, we used the empirical mean of the stopping times that were smaller than $n_{\max}$ on the sample that we obtained by simulation, with 20% of the stopping times being $n_{\max}$ itself. In the 'conditional mean' column, we plotted an estimate of $\tau' \mid \tau' < n_{\max}$, i.e. the stopping time given that we stop early (and hence reject the null).

For comparison, we also show the number of events that one would need under the Gaussian non-sequential approximation of Schoenfeld (1981), and under the continuous monitoring version of the O'Brien-Fleming procedure. In order to judge Schoenfeld's approximation, we report the number of events required to achieve 80% power. This is equivalent to treating the logrank statistic as if it were normally distributed, and rejecting the null hypothesis using a z-test for a fixed number of events. The power analysis of this procedure is classical, and the number of events required is $n^S_{\max} = 4(z_\alpha + z_\beta)^2 / \log^2 \theta_1$, where z_α, and z_β are the α, and β-quantiles of the standard normal distribution. In the case of the continuous monitoring version of O'Brien-Fleming's procedure, we estimated the number of events $n^{OF}_{\max}$ needed to achieve 80% as follows. For each experimental setting (m_0, m_1, θ), we generated 10^4 realizations of the survival process under the alternative hypothesis under consideration and computed the corresponding trajectories of the logrank statistic. For each possible value n of $n^{OF}_{\max}$, we computed the fraction of trajectories for which O'Brien-Fleming's procedure correctly stopped when used with the maximum number of events set to n. We report as an estimate of the true $n^{OF}_{\max}$ the first value of n for which this fraction is higher than 80%, our predefined power.

3 | Accumulation Bias

Abstract

Studies accumulate over time and meta-analyses are mainly retrospective. These two characteristics introduce dependencies between the *analysis time*, at which a series of studies is up for meta-analysis, and results within the series. Dependencies introduce bias – *Accumulation Bias* – and invalidate the sampling distribution assumed for p-value tests, thus inflating type-I errors. But dependencies are also inevitable, since for science to accumulate efficiently, new research needs to be informed by past results. Here, we investigate various ways in which *time* influences error control in meta-analysis testing. We introduce an *Accumulation Bias Framework* that 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 affect p-value-based tests in a similar manner. This leads to two main conclusions. First, Accumulation Bias is inevitable, and even if it can be approximated and accounted for, no valid p-value tests can be constructed. Second, tests based on likelihood ratios withstand Accumulation Bias: they provide bounds on error probabilities that remain valid despite the bias. We leave the reader with a choice between two proposals to consider *time* in error control: either treat individual (primary) studies and meta-analyses as two separate worlds – each with their own timing – or integrate individual studies in the meta-analysis world. Taking up likelihood ratios in either approach allows for valid tests that relate well to the accumulating nature of scientific knowledge. Likelihood ratios can be interpreted as betting profits, earned in previous studies and invested in new ones, while the meta-analyst is allowed to cash out at any time and advise against future studies.

Introduction

> Meta-analysis refers to the statistical synthesis of results from a series of studies. [...] the synthesis will be meaningful only if the studies have been collected systematically. [...] The formulas used in meta-analysis are extensions of formulas used in primary studies, and are used to address similar kinds of questions to those addressed in primary studies.
>
> –Borenstein, Hedges, Higgins & Rothstein (2009, pp. xxi-xxiii)

> To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.
>
> –Fisher (1938, p. 18)

These two quotes conflict. Most meta-analyses are retrospective and consider the number of studies available – after the literature has been searched systematically – as a given for the statistical analysis. *P*-value based statistical tests, however, are intended to be prospective and require the sample size – or the stopping rule that produces the sample – to be set specifically for the planned statistical analysis. The second quote, by the *p*-value's popularizer Ronald Fisher, is about primary studies. But this prospective rationale influences meta-analysis as well because it also involves the size of the study series: *p*-value tests assume that the number of studies – so the timing of the meta-analysis – is predetermined or at least unrelated to the study results. So by using *p*-value methods, conventional meta-analysis implicitly assumes that promising initial results are just as likely to develop into (large) series of studies as their disappointing counterparts. Conclusive studies should just as likely trigger meta-analyses as inconclusive ones. And so the use of *p*-value tests suggests that results of earlier studies should be unknown when planning new studies as well as when planning meta-analyses. Such assumptions are unrealistic and actively argued against by the *Evidence-Based Research Network* (Lund et al., 2016) part of the movement to reduce research waste (Chalmers and Glasziou, 2009; Chalmers et al., 2014). But ignoring these assumptions invalidates conventional *p*-value tests and inflates type-I errors.

P-values are based on tail areas of a test statistic's sampling distribution under the null hypothesis, and thus require this distribution to be fully specified. In this chapter we show that the standard normal *Z*-distribution generally assumed (e.g. Borenstein et al. (2009)) is not an appropriate sampling distribution. Moreover, we believe that no sampling distribution can be specified that fully represents the variety of processes in accumulating scientific knowledge and all decisions made along the way. We need a more flexible approach to testing that controls errors regardless of the process that spurs the meta-analysis.

When dependencies arise between study series size or meta-analysis timing and results within the series, bias is introduced in the estimates. This bias is inherent to accumulating data, which is why we gave it the name *Accumulation Bias*. Various forms of Accumulation Bias have been characterized before, in very general terms as "bias introduced by the order in which studies are conducted" (Whitehead, 2002, p. 197) and more specif-

ically, such as bias caused by the dependence of follow-up studies on previous studies' significance and the dependence of meta-analysis timing on previous study results (Ellis and Stewart, 2009). Also, more elaborate relations were studied between the existence of follow-up studies, study design and meta-analysis estimates (Kulinskaya et al., 2016). Yet no approach to confront these biases has been proposed.

In this chapter we define *Accumulation Bias* to encompass processes that not only affect parameter estimates but also the shape of the sampling distribution, which is why only approximation and correction for bias does not achieve valid p-value tests. We illustrate this by an example in Section 3.2, right after we give a general introduction to Accumulation Bias in Section 3.1 with its relation to publication bias (Section 3.1.1) and an informal characterization of the direction of the bias (Section 3.1.2). By presenting its diversity, we argue throughout the chapter that any efficient scientific process will introduce some form of Accumulation Bias and that the exact process can never be fully known. We collect the various forms of Accumulation Bias into one framework (Section 3.3) and show that all are related to the *time* aspect in meta-analysis. The framework incorporates dependencies mentioned by Whitehead (2002), Ellis and Stewart (2009) and Kulinskaya et al. (2016) as well the effect of multiple testing over time in living systematic reviews Simmonds et al. (2017). We conclude that some version of these biases will also be introduced by *Evidence-Based Research*.

Our framework specifies *analysis time probabilities* – with behavior familiar from survival analysis – and distinguishes two approaches to error control: conditional on time (Section 3.4.1) and surviving over time (Section 3.4.2). We show that general meta-analyses take the former approach, while existing methods for living systematic reviews take the latter. However, neither of the two is able to analyze study series affected by partially unknown processes of Accumulation Bias (Section 3.4.3). After an intermezzo on evidence that indeed such processes are already at play in Section 3.5, we introduce a general form of a test statistic that is able to withstand any Accumulation Bias process: the likelihood ratio. We specify bounds on error probabilities that are valid despite the existing bias, for error control conditional on time (Section 3.6.1) as well as surviving over time (Section 3.6.2). The reader is left to choose between the two; the consequences of either preference are specified in Section 3.7. We try to give intuition on why both are still possible in Section 3.6.1 and Section 3.6.2 respectively, but also give some extra intuition on the magic of likelihood ratios in Section 3.8: Likelihood ratios have an interpretation as betting profit that can be reinvested in future studies. At the same time, the meta-analyst is allowed to cash out at any time and advise against future studies. Hence, the likelihood ratio relates the statistics of Accumulation Bias to the accumulating nature of scientific knowledge, which is critical in reducing research waste.

3.1 Accumulation Bias

Any meta-analyst carries out a meta-analysis under the assumption that synthesizing previous studies will add to what is already known from existing studies. So meta-analyses are mainly performed when the series of studies has reached a meaningful size. What is considered meaningful varies considerably: 16 and 15 studies per meta-analysis were

reported to be the median numbers in *Medline* meta-analyses from 2004 and 2014 (Moher et al., 2007a; Page et al., 2016), while 3 studies per meta-analysis were reported in *Cochrane* meta-analyses from 2008 (*Cochrane Database of Systematic Reviews* Davey et al. (2011)). Since meta-analyses are performed on research hypotheses that have spurred a certain study series size, they always report estimates that are conditioned on the availability of such a series. The crucial point is that not all pilot studies or small study series will reach a meaningful size. Which ones do might depend on results in the series. Apart from the dependent size of the study series, the exact timing of a meta-analysis can also depend on the available results. The completion of a highly powered or otherwise conclusive study, for example, might be considered to finalize the series and trigger a meta-analysis. So meta-analyses also report estimates conditioned on the consideration that a systematic synthesis will be informative. Both dependencies – series size and meta-analysis timing – introduce bias: Accumulation Bias.

3.1.1 Accumulation Bias vs. publication bias

Publication bias refers to the practice that studies with nonsignificant, or more general, unsatisfactory results have smaller probability to be published than studies with significant, satisfactory results. So unsatisfactory studies are performed, but do not reach the meta-analyst because they are stashed away in a file drawer (Rosenthal, 1979). Accumulation Bias, on the other hand, refers to some studies or meta-analyses not being performed at all, as a result of previous findings in a series of studies. In a file drawer-free world, Accumulation Bias would still exist. But Accumulation Bias is a manageable problem because it does not operate at the individual study level. Conditional on the fact that a second study is performed, the second study is an unbiased sample. Conditional on the fact that a third study is performed, for whatever reason, the third study is an unbiased sample. So bias is introduced at the level of the series, not at the study level. This is different for publication bias, where, conditional on being published, the studies available are not an unbiased sample. We exploit the difference in this chapter by considering *time* in error control.

Of course, Accumulation Bias and publication bias are not alone in their effects on meta-analysis reporting. All sorts of *significance chasing biases* – selective-outcome bias, selective analysis reporting bias and fabrication bias – might be present in the study series up for meta-analysis, and can lead to "wrong and misleading answers" (Ioannidis, 2010, p. 169). But for a world in which these biases are overcome, we also need tests that reflect how scientific knowledge accumulates.

3.1.2 Accumulation Bias' direction

Accumulation Bias in estimates is mainly bias in the satisfactory direction, which means that the effect under study is overestimated. This is the case for bias caused by the size of the studies series when (overly) optimistic initial estimates (either in individual studies or in intermediate meta-analyses) give rise to more studies, while disappointing results terminate a series of studies. This is also the case when the timing of the meta-analysis is based on an (overly) optimistic last study estimate or an (overly) optimistic meta-analysis

synthesis is considered the final one. We focus on this satisfactory direction of Accumulation Bias and will only briefly discuss other possibilities in Section 3.4.3 and Section 3.5.1. We introduce the wide variety of possible dependencies in an *Accumulation Bias Framework* in Section 3.3, which has a generality that also includes Accumulation Bias without a clear direction. But we first present Accumulation Bias' effects on error control by an example.

3.2 A *Gold Rush* example: new studies after finding significant results

We study the effect of Accumulation Bias by a simple example. Its simplicity allows us to calculate the exact amount of bias in the test statistic and investigate the additional effect on the sampling distribution. The example given in this section is an extension of the toy example introduced by Ellis and Stewart (2009). We call this example by *Gold Rush* because it describes how new studies go looking for more results after finding initial statistical significance. In the current culture of scientific practice, statistical significance can be seen as the currency of scientific success. After all, significant results achieve the future possibility to pay off in publications, grants and tenure positions. When a gold rush for statistical significance presents itself in a series of studies, dependencies arise between the size of the series and the results within: Accumulation Bias. We specify this mechanism in detail in Section 3.2.2 and Section 3.2.3, after we simplified our meta-analysis setting to common/fixed-effects meta-analysis in Section 3.2.1. We present the resulting bias in the test estimates in Section 3.2.4 and its additional effects on the sampling distribution and testing in Section 3.2.5 and Section 3.2.6. In Section 3.2.7 we conclude by pointing out the very mild condition needed for some form of *Gold Rush* Accumulation Bias to occur

3.2.1 Common/fixed-effect meta-analysis

This chapter discusses meta-analysis in its simplest form, which is common-effect meta-analysis, also known as fixed-effect meta-analysis. This restriction does not mean that more complex forms of meta-analysis, such as random-effects meta-analysis and meta-regression, do not suffer from the problems mentioned in this chapter. The reason for simplification is to reduce the complexity in quantifying the problem, part of showing that quantification is not enough. For an example of Accumulation Bias in random-effects estimates we refer to Kulinskaya et al. (2016).

Common-effect meta-analysis derives a combined Z-score from the summary statistics of the available studies. This combined Z-score is used as a test statistic in two-sided meta-analysis testing by comparing it to the tails of a standard normal distribution. This is equivalent to assessing whether its absolute value is more than $z_{\frac{\alpha}{2}}$ standard deviations away from zero (larger than 1.960 for $\alpha = 0.05$). We simplify the setting by assuming studies with equal standard deviations to obtain an easy to handle expression for the combined Z-score of t available studies. We denote this meta-analysis Z-score by $Z^{(t)}$ and derive it as the weighted average over the study Z-scores $Z_1, \ldots, Z_t$, shown in its general

form in (3.1a) and in (3.1b) under the assumption of equal study sizes:

$$Z^{(t)} = \frac{\sum_{i=1}^{t} \sqrt{n_i} Z_i}{\sqrt{N^{(t)}}} \quad \text{with} \quad N^{(t)} = \sum_{i=1}^{t} n_i \tag{3.1a}$$

$$= \frac{1}{\sqrt{t}} \sum_{i=1}^{t} Z_i \quad (n_1 = n_2 = \cdots = n_t = n). \tag{3.1b}$$

See Appendix Section 3.A for a derivation from the mean difference notation in Borenstein et al. (2009).

3.2.2 *Gold Rush* new study probabilities

In our *Gold Rush* example, we assume the following dependency within a series of studies: each study in a series has a larger probability to be replicated – and therefore expanding the series of studies – if the study shows a significant positive effect. So the existence of a new study is dependent on the significance and sign of the results of its predecessor.

T is the random variable that denotes the maximum size of a study series – the time at which the search stops. We enumerate time by the order of appearance in a study series, with $t = 1$ for the pilot study, $t = 2$ for the second study (so now we have a two-study series) etc. So we use t to denote the number of studies available for meta-analysis at any time point: our notion of time is not related to actual dates at which studies are performed. The maximum time T is usually unknown since more studies might be performed in the future. $T \geq 2$ means that the series has not halted after the first initial study, but that it is unknown how many replications will eventually be performed. In our extended *Gold Rush* example, we present the Accumulation Bias process by the probability that the maximum size is at least one study larger than the current size ($T \geq t + 1$), and do so using six parameters. We denote these parameters by the *new study probabilities*, since they indicate the probability that a follow-up study is performed when the result of the current study is available:

$$\begin{aligned}
\omega_{\text{S}}^{(1)} &:= \mathbf{P}\left[T \geq 2 \,\middle|\, T \geq 1, Z_1 \geq z_{\frac{\alpha}{2}} \right] &&= 1 \\
\omega_{\text{X}}^{(1)} &:= \mathbf{P}\left[T \geq 2 \,\middle|\, T \geq 1, Z_1 \leq -z_{\frac{\alpha}{2}} \right] &&= 0 \\
\omega_{\text{NS}}^{(1)} &:= \mathbf{P}\left[T \geq 2 \,\middle|\, T \geq 1, |Z_1| < z_{\frac{\alpha}{2}} \right] &&= 0.1,
\end{aligned}$$

$$\text{for all } t \geq 2: \tag{3.2}$$

$$\begin{aligned}
\omega_{\text{S}}^{(t)} = \ \omega_{\text{S}} &:= \mathbf{P}\left[T \geq t+1 \,\middle|\, T \geq t, Z_t \geq z_{\frac{\alpha}{2}} \right] &&= 1 \\
\omega_{\text{X}}^{(t)} = \ \omega_{\text{X}} &:= \mathbf{P}\left[T \geq t+1 \,\middle|\, T \geq t, Z_t \leq -z_{\frac{\alpha}{2}} \right] &&= 0 \\
\omega_{\text{NS}}^{(t)} = \omega_{\text{NS}} &:= \mathbf{P}\left[T \geq t+1 \,\middle|\, T \geq t, |Z_t| < z_{\frac{\alpha}{2}} \right] &&= 0.02.
\end{aligned}$$

We distinguish between the influence of the first (pilot) study ($\omega_S^{(1)}$, $\omega_X^{(1)}$ and $\omega_{NS}^{(1)}$) and the others (ω_S, ω_X and ω_{NS}) since pilot studies are carried out with future studies in mind, and therefore replications have higher probability after the first than after other studies in the series, also in case the pilot study is not significant. We assume that no new study is performed when a significant negative result is obtained ($\omega_X^{(1)} = \omega_X = 0$) and new studies are always performed after positive significant findings, the satisfactory result ($\omega_S^{(1)} = \omega_S = 1$). Nonsignificant results have a small, but not negligible probability to spur new studies ($\omega_{NS}^{(1)} = 0.1$, $\omega_{NS} = 0.02$).

3.2.3 *Gold Rush* new study probabilities' independence from data-generating hypothesis

In the following we use $\mathbf{P}_1$ to express probabilities under the alternative hypothesis and $\mathbf{P}_0$ to express probabilities under the null hypothesis. Our new study probabilities in (3.2) were given without reference to any of these hypotheses, to make explicit that they depend solely on the data (or summary statistic Z_t) and the behavior of the researchers; not on the hypothesis that generated the data. So $\mathbf{P}$ in these definitions can be read as $\mathbf{P}_1$ as well as $\mathbf{P}_0$.

In the next sections we focus on *Gold Rush* Accumulation Bias under the null hypothesis and its effect on type-I error control. The values in rightmost column of (3.2) are introduced to obtain estimates for the Accumulation Bias in the test estimates. These values are not supposed to be realistic, but are chosen to demonstrate the effect of Accumulation Bias as clearly as possible. The extreme values 1 for $\omega_S^{(1)}$ and ω_S given in (3.2) support the simulation of large study series under the null hypothesis. The small values for $\omega_{NS}^{(1)}$ and ω_{NS} are chosen such that the effect of significant findings on the sampling distribution is clearly visible (see Section 3.2.5 and Figure 3.1). For $\alpha = 0.05$, $\omega_S^{(1)} = 1$ implies that, in expectation under the null distribution, all of the 2.5% ($\frac{\alpha}{2}$) positively significant pilot studies under the null hypothesis become a two-study series, while $\omega_{NS}^{(1)} = 0.1$ indicates that, since an expected 95% ($1-\alpha$) of pilot studies is not significant under the null hypothesis, 9.5% ($0.1 \cdot 95\%$) become a two-study series. For study series beyond the pilot study and its replication, this setup entails that in all studies, except for the last and the first, the fraction of significant findings is more than half, since $\omega_S = 0.02$ implies that only $0.02 \cdot 95\% = 1.9\%$ nonsignificant studies grow into a larger study series: the expected fraction of significant studies in growing series under the null hypothesis converges to $2.5/(2.5 + 1.9) = 0.6$: the conditional probability of getting a positive finding conditional on another study being done after the first pilot (pilot) study.

3.2.4 *Gold Rush* Accumulation Bias' estimates under the null hypothesis

The new study probability parameters in (3.2) are much larger when results are positively significant than when they are not. As a result, it occurs more often that study series have many significant studies than that they have only a few. While the expectation of a Z-score is 0 under the null hypothesis for each individual study (for all t: $\mathbf{E}_0[Z_t] = 0$), the expectation of a study that is part of a series of studies is larger. This shift in expectation

introduces the Accumulation Bias in the estimates.

The main ingredient of the bias in the meta-analysis $Z^{(t)}$-score is the bias in the individual study Z_t-scores, conditional on being part of a series. This is already apparent for the pilot study, which we use as an example by expressing its expected value under the null hypothesis, given that it has a successor study: $\mathbf{E}_0[Z_1 \mid T \geq 2]$. This conditional expectation is a weighted average of two other expectations that are conditioned further based on the events that lead to a new study according to (3.2): $\mathbf{E}_0\left[Z_1 \,\middle|\, Z_1 \geq z_{\frac{\alpha}{2}}\right]$, Z_1 from the right tail of the null distribution, and the nonsignificant results with expectation $\mathbf{E}_0\left[Z_1 \,\middle|\, |Z_1| < z_{\frac{\alpha}{2}}\right]$. We discard negative significant results, since those were given 0 probability to produce replication studies in (3.2). The positive significant and nonsignificant results are weighted by the new study probabilities in (3.2) and the probabilities under the null distribution of sampling from either the tail (α) or the middle part ($1-\alpha$) of the standard normal distribution. A more detailed specification of these components can be found in Appendix Section 3.B. If we assume a significance threshold of 5% we obtain:

$$\text{for } \alpha = 0.05: \qquad \mathbf{E}_0[Z_1 \mid T \geq 2] = \frac{\int_{z_{\frac{\alpha}{2}}}^{\infty} z \cdot \phi(z)dz \cdot \omega_{\text{S}}^{(1)} \cdot \frac{\alpha}{2} + 0 \cdot \omega_{\text{NS}}^{(1)} \cdot (1-\alpha)}{\omega_{\text{S}}^{(1)} \cdot \frac{\alpha}{2} + \omega_{\text{NS}}^{(1)} \cdot (1-\alpha)} \approx 0.487. \tag{3.3}$$

Here we use the fact that, for $\alpha = 0.05$, $\mathbf{E}_0\left[Z_1 \,\middle|\, Z_1 \geq z_{\frac{\alpha}{2}}\right] = \int_{1.960}^{\infty} z \cdot \phi(z)dz \approx 2.338$, with $\phi()$ the standard normal density function and that $\mathbf{E}_0\left[Z_1 \,\middle|\, |Z_1| < z_{\frac{\alpha}{2}}\right]$ is the expectation of a symmetrically truncated standard normal distribution, which is 0. The value 0.487 is obtained by using the parameter values given in (3.2). For studies in the series later than the pilot study, the expression follows analogously by taking for all $t \geq 2: \omega_{\text{S}}^{(t)} = \omega_{\text{S}}$ and $\omega_{\text{NS}}^{(t)} = \omega_{\text{NS}}$: $\mathbf{E}_0[Z_t \mid T \geq t+1] \approx 1.328$.

To determine the effect on the meta-analysis $Z^{(t)}$-score, we define the expectation under the null hypothesis $\mathbf{E}_0\left[Z^{(t)} \,\middle|\, T \geq t\right]$, conditioned on the availability of a series of size t. To specify this expectation, we use that the last study so-far (for a series of size t, the t^{th} study) is always unbiased since we do not know whether it will spur more studies. After all, in the *Gold Rush* scenario, we assume that the timing of the meta-analysis does not relate to the t^{th} study result, only that the results of the first $t-1$ studies spurred a series of size t and these results are included in the meta-analysis of t studies. As shown in more detail in Appendix Section 3.C, the expression follows from (3.1a) by separately treating the unbiased expectation of 0 and the pilot study. If we assume a significance threshold of 5%, we obtain the general expression in (3.4a) and the expression in (3.4b) under the

Table 3.1. *Expected Z-scores under the null hypothesis in the Gold Rush scenario, under the equal study size assumption, calculated using (3.4b) with $\alpha = 0.05$ and values for $\omega_S^{(1)}$, $\omega_{NS}^{(1)}$, ω_S and ω_{NS} from (3.2). $Z^{(t)}$ is as defined in (3.1b). See Appendix Section 3.G for the code that was used to calculate these values.*

Number of studies (t)	$\mathbf{E}_0[Z_t]$	$\mathbf{E}_0[Z_t \mid T \geq t+1]$	$\mathbf{E}_0\left[Z^{(t)} \middle\vert T \geq t\right]$
1	0.000	0.487	0.000
2	0.000	1.328	0.344
3	0.000	1.328	1.048
4	0.000	1.328	1.572
5	0.000	1.328	2.000
6	0.000	1.328	2.368
7	0.000	1.328	2.695
8	0.000	1.328	2.990
9	0.000	1.328	3.262
10	0.000	1.328	3.515

assumption of equal study sizes ($n_1 = n_2 = \cdots = n_t = n$):

$$\text{for } \alpha = 0.05, \text{ for all } t \geq 2:$$

$$\mathbf{E}_0\left[Z^{(t)} \middle\vert T \geq t\right] \approx \frac{\sqrt{n_1} \cdot 0.487 + \sum_{i=2}^{t-1} \sqrt{n_i} \cdot 1.328 + \sqrt{n_t} \cdot 0}{\sqrt{N^{(t)}}} \tag{3.4a}$$

$$= \frac{0.487 + 1.328(t-2)}{\sqrt{t}}. \tag{3.4b}$$

Table 3.1 shows the Accumulation Bias in the estimates of $\mathbf{E}_0\left[Z^{(t)} \middle\vert T \geq t\right]$ as studies accumulate under the *Gold Rush* scenario, with equal study sizes and values for the new study probabilities given by (3.2).

3.2.5 *Gold Rush* Accumulation Bias' sampling distribution under the null hypothesis

Figure 3.1 shows simulated *Gold Rush* sampling distributions for study series of size two and three in comparison to an individual study Z-distribution. Because the new study probabilities in (3.2) give Z_{t-1}-values below $-z_{\frac{\alpha}{2}}$ zero probability to warrant a successor study, values for the $z^{(t)}$-statistic below $-z_{\frac{\alpha}{2}}$ will be scarce and the larger t is the larger this scarcity will be since only the last study is able to provide such small Z-score estimates. The opposite is the case for values above $z_{\frac{\alpha}{2}}$, which have probability 1 to warrant a new study. As a result, the distribution of the meta-analysis Z-score has negative skew (more mass on the right, more tail to the left). See the comparison to the normal distribution also plotted in Figure 3.1 for a three-study series. Skewness is not the only characteristic that distinguishes the resulting distribution from a standard normal. The variance also

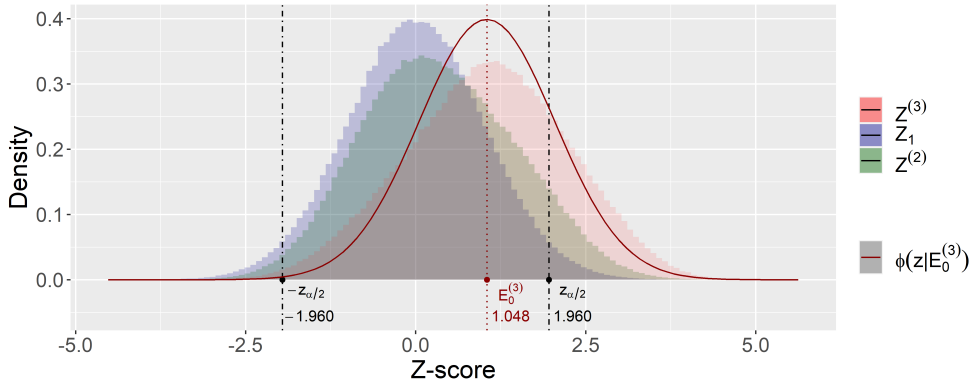

Figure 3.1. *Sampling distributions of meta-analysis $Z^{(t)}$-scores under the null hypothesis in the Gold Rush scenario, under the equal study size assumption, with $\alpha = 0.05$ and values for $\omega_S^{(1)}$, $\omega_{NS}^{(1)}$, ω_S and ω_{NS} from (3.2). $Z^{(t)}$ is as defined in (3.1b). $\phi(z|\mathbf{E}_0{}^{(3)})$ the standard normal density function shifted by $\mathbf{E}_0{}^{(3)}$, with $\mathbf{E}_0{}^{(3)}$ shorthand for $\mathbf{E}_0\left[Z^{(3)} \middle| T \geq 3\right]$. See Appendix Section 3.G for the code that produces the simulation and creation of this figure.*

deviates since the meta-analysis distribution is a mixture distribution. For a two-study meta-analysis $Z^{(2)}$ we obtain a mixture of two conditional distributions, one conditioned on the first study being a significant – sampled from the right tail of the distribution (with probability $\frac{\alpha}{2} \cdot \omega_S^{(1)}$) – and one with the first study nonsignificant – sampled from the symmetrically truncated normal distribution (with probability $(1-\alpha) \cdot \omega_{NS}^{(1)}$). Because the combined distribution on $Z^{(2)}$ is a mixture of the two scenarios, its variance is larger than the variance of either of the two components of the mixture, as we show in Appendix Section 3.D. In Figure 3.1 we see that, with the parameter values from (3.2) the variance of $Z^{(2)}$ and $Z^{(3)}$ are even larger than that of Z_1, even though both $\mathbf{Var}\left\{Z^{(2)} \middle| Z_1 < z_{\frac{\alpha}{2}}\right\}$ and $\mathbf{Var}\left\{Z^{(2)} \middle| |Z_1| \geq z_{\frac{\alpha}{2}}\right\}$ are smaller. Hence the sampling distribution under the null hypothesis of a meta-analysis Z-score deviates from a standard normal under Accumulation Bias due to a non-zero location (the bias), skewness and inflated variance. All three inflate the probability of a type-I error in a standard normal test, as we will study in the next section.

3.2.6 *Gold Rush* Accumulation Bias' influence on p-value tests

Let us now establish the effect of our *Gold Rush* Accumulation Bias on meta-analysis testing when using common/fixed-effects Z-tests. Let $\mathscr{E}_{\text{TYPE-I}}^{(t)}$ indicate the event of a type-I error (significant result under the null hypothesis) in a meta-analysis of t studies and let $\mathbf{P}_0\left[\mathscr{E}_{\text{TYPE-I}}^{(t)} \middle| T \geq t\right] = \mathbf{P}_0\left[|Z^{(t)}| \geq z_{\frac{\alpha}{2}} \middle| T \geq t\right]$ denote the expected rate of type-I errors in a two-sided common/fixed-effect Z-test for studies i up to t conditional on the fact that at least t studies were performed.

We obtain the type-I error rate for this test by simulating the *Gold Rush* scenario, for which

Table 3.2. Inflated type-I error rates for tests affected by bias only ($\widetilde{\mathbf{P}}_0[\mathscr{E}^{(t)}_{\text{TYPE-I}} \mid T \geq t]$) and tests affected by bias as well as impaired sampling distribution ($\mathbf{P}_0[\mathscr{E}^{(t)}_{\text{TYPE-I}} \mid T \geq t]$). Simulated values are under the null hypothesis in the Gold Rush scenario, under the equal study size assumption, with $\alpha = 0.05$ and values for $\omega_S^{(1)}$, $\omega_{NS}^{(1)}$, ω_S and ω_{NS} from (3.2). See Appendix Section 3.G for the code that produces the simulation and creation of this table.

Number of studies (t)	$\widetilde{\mathbf{P}}_0[\mathscr{E}^{(t)}_{\text{TYPE-I}} \mid T \geq t]$	$\mathbf{P}_0[\mathscr{E}^{(t)}_{\text{TYPE-I}} \mid T \geq t]$
2	0.06	0.10
3	0.18	0.23
4	0.35	0.40
5	0.52	0.53

the results are shown in the right hand column of Table 3.2, assuming $\alpha = 0.05$. If only bias would be at play, the sampling distribution under the null hypothesis would be a shifted normal distribution. (3.5) expresses the expected type-I error rate for this bias only scenario, with $\Phi()$ the cumulative normal distribution. The actual inflation in the type-I error rate is larger than shown by this scenario, as illustrated the Table 3.2. The difference between these two type-I error rates for a series of three studies is depicted in Figure 3.1 by the area under the red histogram for $Z^{(3)}$ and the red $\phi(z \mid \mathbf{E}_0{}^{(3)})$ curve below $-z_{\frac{\alpha}{2}}$ and above $z_{\frac{\alpha}{2}}$. We conclude that the effect of Accumulation Bias on testing cannot be corrected by only an approximation of the bias.

$$\widetilde{\mathbf{P}}_0\left[\mathscr{E}^{(t)}_{\text{TYPE-I}} \,\middle|\, T \geq t\right] := 1 - \Phi\left(z_{\frac{\alpha}{2}} - \mathbf{E}_0\left[Z^{(t)} \,\middle|\, T \geq t\right]\right) + \Phi\left(-z_{\frac{\alpha}{2}} - \mathbf{E}_0\left[Z^{(t)} \,\middle|\, T \geq t\right]\right). \quad (3.5)$$

3.2.7 *Gold Rush* Accumulation Bias: When does it occur?

We indicated in Section 3.2.3 that we chose extreme values for parameters $\omega_S^{(1)}$, $\omega_X^{(1)}$, $\omega_{NS}^{(1)}$, ω_S, ω_X and ω_{NS} such that Figure 3.1 would clearly show the bias and distributional change that occurs. However, for any combination of values for which there is a t where $\omega_S^{(t)} \neq \omega_X^{(t)} \neq \omega_{NS}^{(t)}$ Accumulation Bias occurs for series larger than size t and p-value tests that assume a standard normal distribution are invalid.

3.3 The Accumulation Bias Framework

In general, Accumulation Bias in meta-analysis makes the sampling distribution of the meta-analysis Z-score difficult to characterize due to the data dependent size and timing of a study series up for meta-analysis. In this section, we specify both processes in a framework of analysis time probabilities. We use the term *analysis time* because time in meta-analysis is partly based on a *survival time*. A survival time indicates that a subject lives longer than time t (and might still become much older), just as an analysis time indicates that a series up for meta-analysis has at least size t (but might still grow much larger). As such, analysis time probabilities, just as the probabilities in a survival function, do not add up to 1.

Our *Accumulation Bias Framework* uses the following notation for its three key components: $S(t-1)$, $\mathscr{A}^{(t)}$ and $A(t)$. Firstly, $S(t-1)$ can be understood as the survival function in the variable time t that indicates the size of the expanding study series. $S(t-1)$ denotes the probability that the available number of studies is at least t ($\mathbf{P}[T \geq t]$), so the study series has survived past the previous study at $t-1$. Secondly, $\mathscr{A}^{(t)}$ indicates the event that a meta-analysis is performed on a study series of size exactly t. Lastly, $A(t)$ combines the probability that a study series of certain size is available ($S(t-1)$) with the decision $\mathscr{A}^{(t)}$ to perform the analysis on exactly t studies. So the *analysis time probability* $A(t)$ represents the general probability that a meta-analysis of size t – so at *time* t – is performed and is the key to describing the influence of various forms of Accumulation Bias on testing.

3.3.1 Analysis time probabilities

Let $\mathbf{P}\left[\mathscr{A}^{(t)} \,\middle|\, T \geq t, z_1, \ldots, z_t\right]$ denote the probability that a meta-analysis is performed on the first t studies. Just as the *Gold Rush*' new study probabilities from (3.2), this probability can depend on the results in the study series $z_1, \ldots, z_t$. The event $\mathscr{A}^{(t)}$ only occurs if a series of size t is available, so we need to condition on the survival past $t-1$, which can also depend on previous results. When combined, we obtain the following definition[1] of *analysis time probabilities* $A(t)$:

$$\begin{aligned} A(t \,|\, z_1, \ldots, z_t) &:= \mathbf{P}\left[\mathscr{A}^{(t)} \,\middle|\, T \geq t, z_1, \ldots, z_t\right] \cdot S(t-1 \,|\, z_1, \ldots, z_{t-1}), \\ &\text{where we define} \\ S(t-1 \,|\, z_1, \ldots, z_{t-1}) &:= \mathbf{P}[T \geq t \,|\, z_1, \ldots, z_{t-1}]. \end{aligned} \tag{3.6}$$

(3.6) formalizes the idea of analysis time probabilities "depending on previous results" in terms of the individual study Z-scores $z_1, \ldots, z_t$. This is compatible with the Z-test approach in meta-analysis and the dependencies and the *Gold Rush*' new study probabilities that are explicitly expressed in terms of Z-scores. More generally however, in Section 3.3.3 and Section 3.3.4 we extend the definition and allow analysis time probabilities to also depend on the data in the original scale and external parameters.

3.3.2 Analysis time probabilities' independence from the data-generating hypothesis

Just as for the *Gold Rush*' new study probabilities discussed in Section 3.2.2 and Section 3.2.3, the analysis time probabilities $A(t)$ only depend on the data, and are independent from the hypothesis that generated the data. So again, $\mathbf{P}$ in these definitions can

[1] Note that $A(t \mid z_1, \ldots, z_t)$ is defined as a product of two (conditional) probabilities. Calling this product itself a "probability", as we do, can be justified as follows: we currently think of the decision whether to continue studies at time t, i.e. whether $T \geq t$, to be made before the t-th study is performed. But we may also think of the t-study result z_t as being generated irrespective of whether $T \geq t$, but remaining unobserved for ever if $T < t$. If the decision whether $T \geq t$ is made independently of the value z_t, i.e. we add the constraint $\mathbf{P}[T \geq t \,|\, z_1, \ldots, z_{t-1}] = \mathbf{P}[T \geq t \,|\, z_1, \ldots, z_t]$, then the resulting model is mathematically equivalent to ours (in the sense that we obtain exactly the same expressions for $S(t)$, $A(t \mid z_1, \ldots, z_t)$, all error probabilities etc.), but it does allow us to write, by (3.6), that $A(t \,|\, z_1, \ldots, z_t) = \mathbf{P}\left[\mathscr{A}^{(t)}, T \geq t \,\middle|\, z_1, \ldots, z_t\right]$ – so now $A(t \,|\, z_1, \ldots, z_t)$ is indeed a probability.

be read as $\mathbf{P}_1$ as well as $\mathbf{P}_0$. Our definition of $A(t)$ relates to the definition of a *Stopping Rule* by Berger and Berry (1988, pp. 33-34), where they use $x^{(m)}$ to denote a vector of m observations:

> **Definition.** A *stopping rule* is a sequence $\tau = (\tau_0, \tau_1, \dots)$ in which $\tau_0 \in [0,1]$ is a constant and τ_m is a measurable function of $x^{(m)}$ for $m \geq 1$, taking values in $[0,1]$.
>
> τ_0 is the probability of stopping the experiment with no observations (e.g., if it is determined that the experiment is too expensive); $\tau_1(x^{(1)})$ is the probability of stopping after observing the datum $x^{(1)} = x_1$, conditional on having taken the first observation; $\tau_2(x^{(2)})$ is the probability of stopping after observing $x^{(2)} = (x_1, x_2)$, conditional on having taken the first and second observations; etc.

To take the analogy with survival analysis further, we consider the sequence τ defined above by Berger and Berry (1988) to be a sequence of hazards. Instead of using their notation τ we denote the *Stopping Rule* by $\lambda = (\lambda(0), \lambda(1), \dots)$ to emphasize its behavior as a sequence of *hazard functions* and to distinguish time t from the probability $\lambda(t)$ of stopping at that time given that you were able to reach it. The hazard of stopping at time t can depend on previous results and is defined as follows:

$$\lambda(t \,|\, z_1, \dots, z_t) := \mathbf{P}[T = t \,|\, T \geq t, z_1, \dots, z_t]. \tag{3.7}$$

In this chapter we are only interested in cases in which a first study is available, so $\lambda(0) = 0$ (also stated as $\mathbf{P}[T \geq 1] = 1$ in Appendix Section 3.B). The survival $S(t-1)$, the probability of obtaining a series of size at least t (so larger than $t-1$), follows from the hazards by considering that surviving past time $t-1$ means that the series has not stopped at studies i up to and including $t-1$. So for $t \geq 1$:

$$S(t-1 \,|\, z_1, \dots, z_{t-1}) = \prod_{i=0}^{t-1} (1 - \lambda(i \,|\, z_1, \dots, z_i)). \tag{3.8}$$

In many examples, the hazard of stopping at time t, $\lambda(t)$, will depend on the result z_t just obtained. In that case $\lambda(i \,|\, z_1, \dots, z_i) = \lambda(i \,|\, z_i)$ in (3.8) above. But in general $\lambda(t)$ might also depend on some synthesis of all z_i so far. We show some of the variety of forms that $\lambda(t)$, $S(t)$ and $A(t)$ can take in our Accumulation Bias Framework in the following sections.

3.3.3 Accumulation Bias caused by dependent study series size

Our *Gold Rush* example describes an instance of Accumulation Bias that is caused by how the study series size comes about. This is expressed by the $S(t)$ component of the analysis times probability $A(t)$. We represent our *Gold Rush* scenario in terms of our Accumulation Bias framework in next section, followed by variations from the literature that we were able to express in a similar manner.

Gold Rush: dependence on significant study results

The *Gold Rush* scenario operates in a general meta-analysis setting and assumes that there is a single random or prespecified time t at which a study series is up for meta-analysis. This is the approach taken by meta-analyses not explicitly part of a living systematic review. In the *Gold Rush* example the dependency arises in the study series because a t-study series has a larger probability to come into existence when individual study results are significant, and you need a t-study series to perform a t-study meta-analysis. This dependency was characterized by the new study probabilities $\omega_S^{(1)}$, $\omega_{NS}^{(1)}$, ω_S and ω_{NS} from (3.2). The value of $S(t)$, and therefore $A(t)$, can be expressed in terms of these new study probabilities by considering whether $z_1, \ldots, z_{t-1}$ are larger than $z_{\frac{\alpha}{2}}$ (which is 1.960 for $\alpha = 0.05$). Since a meta-analysis is performed only once at a randomly chosen time t, we have $\mathbf{P}[\mathscr{A}^{(t)}] = 1$ for that time t and $\mathbf{P}[\mathscr{A}^{(t)}] = 0$ otherwise. So for the one meta-analysis we obtain:

for t such that $\mathbf{P}[\mathscr{A}^{(t)}] = 1$:

$$A(t \mid z_1, \ldots, z_{t-1}; \alpha) = S(t-1 \mid z_1, \ldots, z_{t-1}; \alpha) = \prod_{i=0}^{t-1} (1 - \lambda(i \mid z_i; \alpha)),$$

$$\text{with } \lambda(0) = 0 \text{ and for all } i \geq 1,\ \lambda(i) \text{ is defined as follows:} \tag{3.9}$$

$$\lambda(i \mid z_i, \alpha) = 1 - \left(\omega_S^{(i)} \cdot \mathbf{1}_{z_i \geq z_{\frac{\alpha}{2}}} + \omega_{NS}^{(i)} \cdot \mathbf{1}_{|z_i| < z_{\frac{\alpha}{2}}} \right)$$

$$\overline{\lambda}_0(i \mid \alpha) := \mathbf{E}_0[\lambda(i \mid Z_i; \alpha)] = 1 - \left(\omega_S^{(i)} \cdot \frac{\alpha}{2} + \omega_{NS}^{(i)} \cdot (1 - \alpha) \right).$$

Therefore, (leaving out the $\lambda(0)$ and summing from $i = 1$ to $t-1$), we obtain the following expressions for the *Gold Rush* analysis time probabilities and its expectations under the null distribution:

$$A(t \mid z_1, \ldots, z_{t-1}; \alpha) = \prod_{i=1}^{t-1} \left(\omega_S^{(i)} \cdot \mathbf{1}_{z_i \geq z_{\frac{\alpha}{2}}} + \omega_{NS}^{(i)} \cdot \mathbf{1}_{|z_i| < z_{\frac{\alpha}{2}}} \right)$$

$$\overline{A}_0(t \mid \alpha) := \mathbf{E}_0[A(t \mid Z_1, \ldots, Z_{t-1}; \alpha)] = \prod_{i=1}^{t-1} \left(\omega_S^{(i)} \cdot \frac{\alpha}{2} + \omega_{NS}^{(i)} \cdot (1 - \alpha) \right). \tag{3.10}$$

Kulinskaya et al. (2016): dependence on meta-analysis estimates

Kulinskaya et al. (2016) report biases that result from dependencies between a current meta-analysis estimate and the decision to perform a new study. Since their focus is on bias, they do not discuss issues of multiple testing over time, which would arise if their cumulative meta-analyses estimates were tested. In this section we assume that the timing of the meta-analysis test is independent from the estimates that determined the accumulation of the series to its current size, as if a test were done by a second unknowing meta-analyst. This scenario is hinted at by Kulinskaya et al. (2016, p. 296) in the statement "When a practitioner or a meta-analyst finds several trials in the literature, a particular decision-making scenario may have already taken place." We postpone the discussion of multiple testing to Equation 3.3.3. In this estimation setting, the decision to perform new

studies is determined not by the meta-analysis Z-scores $Z^{(t-1)}$, but by the meta-analysis estimates on the original scale $M^{(t-1)}$ (notation adopted from Borenstein et al. (2009), see Appendix Section 3.A), in relation to a minimally clinically relevant effect Δ^{H1}. A minimally clinically relevant effect is the effect that should be used to power a trial (in the alternative distribution H1), and therefore, the effect that the researchers of the study do not want to miss. Kulinskaya et al. (2016) consider three models for the study series accumulation process: the *power-law model* and the *extreme-value model* and the *probit model*. The models relate the probability of a new study to the cumulative meta-analysis estimate of the study series so far and are inspired by models for publication bias. Although all three models can be recast in our framework, we demonstrate this only for the power law model that uses one extra parameter τ to relate the previous meta-analysis estimate $M_{(t-1)}$ to $S(t)$. Just as in the *Gold Rush* scenario, we must assume that a meta-analysis test is performed only once at a randomly chosen time t. So only at that time t $\mathbf{P}[\mathscr{A}^{(t)}] = 1$ and $\mathbf{P}[\mathscr{A}^{(t)}] = 0$ otherwise. We obtain the following expression for the Kulinskaya et al. (2016) *power-law model*:

for t such that $\mathbf{P}[\mathscr{A}^{(t)}] = 1$:

$$A\left(t \,\middle|\, M^{(t-1)}; \Delta^{H1}, \tau\right) = S\left(t-1 \,\middle|\, M^{(t-1)}; \Delta^{H1}, \tau\right) = \prod_{i=0}^{t-1} \left[1 - \lambda\left(i \,\middle|\, M^{(t-1)}; \Delta^{H1}, \tau\right)\right], \tag{3.11}$$

with $\lambda(0) = \lambda(1) = 0$, and for all $i \geq 2$, $\lambda(i)$ is defined as follows:

$$\lambda\left(i \,\middle|\, M^{(i-1)}; \Delta^{H1}, \tau\right) = 1 - \left(\frac{M^{(i-1)}}{\Delta^{H1}}\right)^{\tau}, \tag{3.12}$$

for $0 < M^{(i-1)} < \Delta^{H1}$ and 1 (so $1 - \lambda = 0$) otherwise.

According to this model, no further studies are performed as soon as an estimate as large as Δ^{H1} is found. For estimates smaller than Δ^{H1}, the closer the estimate is to Δ^{H1}, the larger the probability of a subsequent study. Just as in the *Gold Rush* example, this model will introduce bias as well as skew the sampling distribution of the data under the null hypothesis since initial studies with large estimates have larger probability to end up in study series of considerable size than small initial estimates do. When the initial study gives a large overestimation of the effect, this overestimation stays present in the subsequent meta-analysis estimates and keeps influencing the probability of subsequent studies. Therefore, this model shows the effect of early studies in the series even more clearly than the *Gold Rush* example. However, the accumulation bias does have a cap, since estimates larger than Δ^{H1} do not introduce new replication studies.

Whitehead (2002): dependence on early study results

> Bias may also be introduced by the order in which studies are conducted. For example, large-scale clinical trials for a new treatment are often undertaken following promising results from small trials. [...] given that a meta-analysis

> is being undertaken, larger estimates of treatment difference are more likely from the small early studies than from the later larger studies.
>
> –Whitehead (2002, p. 197)

Whitehead (2002) mentions a dependence between the results of the small early studies in a series and the size of the series. This influence could either be based on the significance of early findings, such as in the *Gold Rush* example (Section 3.3.3), or on the estimates in the initial studies, such as in the power law model from Kulinskaya et al. (2016) (Equation 3.3.3). Whitehead (2002) does not give sufficient details to specify this dependency explicitly, but we are confident that it will fit in our Accumulation Bias framework.

Two ways to approach this Accumulation Bias are given in Whitehead (2002). The first is to exclude early studies from the meta-analyses, either in the main analysis or in a sensitivity analysis. The second way is to ignore the problem, since the small studies will have little effect on the overall estimate. In Section 3.6 we show that any small initial study dependency that can be expressed in terms of $A(t)$ can be dealt with by tests using likelihood ratios.

Living Systematic Reviews: dependence on significant meta-analyses + multiple testing

> A living systematic review (LSR) should keep the review current as new research evidence emerges. Any meta-analyses included in the review will also need updating as new material is identified. If the aim of the review is solely to present the best current evidence standard meta-analysis may be sufficient, provided reviewers are aware that results may change at later updates. If the review is used in a decision-making context, more caution may be needed. When using standard meta-analysis methods, the chance of incorrectly concluding that any updated meta-analysis is statistically significant when there is no effect (the type I error) increases rapidly as more updates are performed.
>
> –Simmonds, Salanti, McKenzie & Elliott (2017, p. 39)

In living systematic reviews, the aim is to have a meta-analysis available to present the current evidence, thus synthesizing the t studies available at a certain time. The current meta-analysis estimate might be used to decide whether further studies should be performed. In that case $S(t-1)$, the probability that a study series of size t is available – so that a study series has expanded beyond series size $t-1$ – depends on the meta-analysis estimate $Z^{(t-1)}$ at the previous study's meta-analysis. Because the review is continuously updated, $\mathbf{P}[\mathscr{A}]$ is always 1, and living systematic reviews can be described by the following analysis time probability $A(t)$:

$$\begin{aligned} A\left(t \,\middle|\, z^{(1)},\ldots,z^{(t)};z_{\frac{\alpha}{2}}\right) &= \mathbf{P}\left[\mathscr{A}^{(t)} \,\middle|\, T \geq t\right] \cdot S\left(t-1 \,\middle|\, z^{(1)},\ldots,z^{(t)};z_{\frac{\alpha}{2}}\right) \\ &= S\left(t-1 \,\middle|\, z^{(1)},\ldots,z^{(t-1)};z_{\frac{\alpha}{2}}\right) = \prod_{i=0}^{t-1}(1-\lambda\left(i \,\middle|\, z^{(i)};z_{\frac{\alpha}{2}}\right)). \end{aligned} \tag{3.13}$$

The quote above warns against decisions based on the continuously updated meta-analysis using a fixed threshold $z_{\frac{\alpha}{2}}$. Living systematic reviews experience multiple testing problems of a kind that are familiar from statistical monitoring of individual clinical trials (Proschan et al., 2006). If the study series is stopped as soon as a significance threshold is reached, and the obtained meta-analysis is considered the final one, then this final meta-analysis test has an increased chance of a type-I error. So the warning is not to use the following simple stopping rule:

$$\lambda\left(i \,\middle|\, z^{(i)}; z_{\frac{\alpha}{2}}\right) = \mathbf{1}_{|Z^{(i)}| \geq z_{\frac{\alpha}{2}}}. \tag{3.14}$$

Various corrections to significance thresholds are proposed that relate intermediate looks to a maximum sample size or information size. These corrected thresholds depend on α and the fraction of sample size or information size available at time t. Examples of such methods are *Trial sequential analysis* (Brok et al., 2008; Wetterslev et al., 2008) and *Sequential meta-analysis* (Whitehead, 2002, Ch. 12) (Whitehead, 1997; Higgins et al., 2011). For an overview see Simmonds et al. (2017). In general, (3.13) and (3.14) show that any dependency between "the best current evidence" and the accumulation of future studies is part of our Accumulation Bias Framework. We discuss the approach to error control taken by the corrected thresholds in Section 3.4.2.

3.3.4 Accumulation Bias caused by dependent meta-analysis timing

We described various forms of Accumulation Bias that are caused by how the study series size comes about, but dependencies are also introduced by how the meta-analysis itself arises. This is expressed by the $\mathbf{P}\left[\mathscr{A}^{(t)}\right]$ component of the analysis times probabilities $A(t)$. We only found one such process mentioned in the literature and will discuss it in the next section.

Ellis and Stewart (2009): dependence on the right amount of positive findings

> Meta-analysis times are subtle. A train of negative findings would generally not stimulate a meta-analysis. Nor would a string of very positive findings. [...] All this makes the analysis of explicitly defined meta-analysis times very difficult. We conclude that study of bias in meta-analysis based on parametric modeling of meta-analysis times is problematical.
>
> –Ellis & Stewart (2009, pp. 2454-2455)

Ellis and Stewart (2009) do not give an explicit model that we can interpret in terms of $A(t)$, but indicate that it should depend on the study findings Z_i, or in the original scale, $\overline{D}_i$ (notation adapted from Borenstein et al. (2009), see Appendix Section 3.A). Given the quote above, the amount of very positive findings should not be too large, and not too small. Though exact parametric modeling indeed stays problematical, we can assume that a positive finding is a study estimate larger than the minimally clinically relevant effect Δ^{H1}, define the right amount of positive findings to be in the region [a, b], and show that

this fits in our Accumulation Bias Framework by expressing a possible model for $A(t)$:

$$\begin{aligned}
&\text{for } t \text{ such that } S(t-1) = 1: \\
&\quad A\left(t \,\middle|\, \overline{D}_1, \ldots, \overline{D}_t; a, b\right) = \mathbf{P}\left[\mathscr{A}^{(t)} \,\middle|\, T \geq t, \overline{D}_1, \ldots, \overline{D}_t; a, b\right] \cdot S\left(t-1 \,\middle|\, \overline{D}_1, \ldots, \overline{D}_{t-1}; a, b\right) \\
&\qquad\qquad = \mathbf{P}\left[\mathscr{A}^{(t)} \,\middle|\, T \geq t, \overline{D}_1, \ldots, \overline{D}_t; a, b\right] = \mathbf{1}_{C \in [a,b]} \\
&\qquad \text{with} \quad C = \sum_{i=1}^{t} \mathbf{1}_{\overline{D}_i > \Delta^{H1}}.
\end{aligned} \tag{3.15}$$

3.3.5 Accumulation Bias caused by Evidence-Based Research

> New research should not be done unless, at the time it is initiated, the questions it proposes to address cannot be answered satisfactorily with existing evidence.
>
> –Chalmers & Glasziou (2009)

In 2009, the term *Research Waste* was coined and this key recommendation was made. The recommendation further specifies that existing evidence should be obtained by a systematic review and summarized with a meta-analysis. But how exactly to answer the question whether new research is necessary or wasteful remained unclear. Nevertheless, the recommendation was important enough to be repeated, as was first done in an entire series on Research Waste with a specific recommendation on setting research priorities Chalmers et al. (2014) and later in a paper that gave the recommendation its official name: *Evidence-Based Research* Lund et al. (2016). Support for these recommendations was provided by various retrospective cumulative meta-analyses that show how many studies were still performed while satisfactory evidence was already available. These cumulative meta-analyses judge "satisfactory evidence" based on a significance threshold, usually uncorrected for multiple testing (e.g. Fergusson et al. (2005)), which reminds us of the Accumulation Bias that occurs in living systematic reviews (Equation 3.3.3).

The larger consequence, however, is that Accumulation Bias is caused by any dependencies between results and series size and meta-analysis timing, and that Evidence-Based Research introduces such dependencies. Inspecting previous results to decide whether new research is necessary or wasteful therefore always introduces Accumulation Bias, whether it based on uncorrected or corrected thresholds. Also more subtle decision methods – implicit rather than based on thresholds – introduce Accumulation Bias, as was shown by Kulinskaya et al. (2016). In fact, they describe the rationale behind their models – among which the *power-law* model (Equation 3.3.3) – as an example of bias introduced by guidelines to decide on "the usefulness of a new study" "with direct reference to existing meta-analysis." Kulinskaya et al. (2016, p. 297).

So Evidence-Based Research causes bias, and our Accumulation Bias Framework demonstrates how it might affect the sampling distribution, whether based on explicit thresholds or implicit decision making. Does this mean that we cannot make Evidence-Based Research decisions to avoid research waste, while also controlling type-I errors? Fortunately,

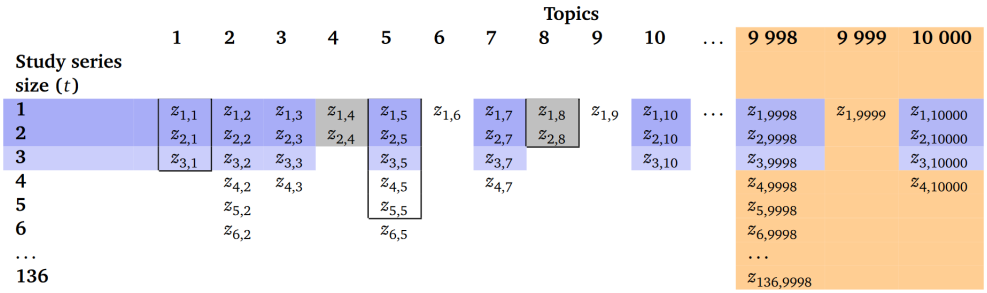

Figure 3.2. *Possible 2001 state of a database of study series per topic, visualizing what study series are taken into account in the two approaches to error control: conditional on time (blue and grey) and surviving over time (orange).*

we do not need to be that pessimistic and can still embrace *Evidence-Based Research*. In Section 3.6 we show that tests based on likelihood ratios withstand Accumulation Bias and are very well suited to reduce research waste. But to do so, we first need to specify exactly what role is played by *time* in error control.

3.4 *Time* in error control

Over time new study series are initiated, studies are added to existing study series and more meta-analyses are performed. To visualize how this process relates to error control, we need to start with a specific state of this expanding system. In 2001 an estimated minimum of 10 000 medical topics were covered in over half a million studies, thus requiring 10 000 meta-analyses if all were synthesized in a database such as the *Cochrane Database of Systematic Reviews* Mallett and Clarke (2003). The number of studies in a series varied between 2 and 136, which we can use to describe the 2001 state of a possible database, that to be complete, also includes many unreplicated pilot studies. We could visualize this database in a table, with studies in the rows, topics in the columns and many missing entries. A sketch is shown in Figure 3.2.

The conventional approach to error control, which we used to show the influence of *Gold Rush* Accumulation Bias in meta-analysis testing in Section 3.2.6, is a conditional approach. Since conventional meta-analysis does not raise any multiple testing issues, there is a hidden assumption that the timing of a meta-analysis $\mathscr{A}^{(t)}$ is independent from the data and each study series experiences only one meta-analysis. In Section 3.3.3 we took the t at which the sole meta-analysis is conducted to be either random or prespecified. This is shown in Figure 3.2 by the black box enclosing the available studies on Topic 1. Other possible study series up for meta-analysis are shown by the boxes enclosing studies on Topic 5 and 8. Note that by assuming only one meta-analysis, a study series might continue growing but not be fully analyzed, as shown for Topic 5.

In the conditional approach to error control, a three-study series (Z_1, Z_2, Z_3) produces a possible draw from the $Z^{(3)}$ sampling distribution. If we test our draw, the type-I error rate

is defined as the fraction of t-study series that is considered significant if all t-study series were to be sampled from the null distribution. The question is: What study series are taken into account to specify this fraction? This is visualized in Figure 3.2 by the dark blue and grey shading for $t = 2$ and the dark blue and lighter blue shading for $t = 3$. The unshaded topics and change of color between $t = 2$ and $t = 3$ show the flaw of this approach: some series might not survive up until a specific time t, as for instance shown by the grey studies that are part of $t = 2$ but not part of the error control for $t = 3$. We also do not want every series to survive up until any arbitrary time t to avoid research waste (Chalmers and Glasziou, 2009). The crucial point is that the series that do survive are no random sample from all possible t-study series. This is another illustration of Accumulation Bias such as the *Gold Rush* scenario. The series deviates even more from the assumption of a random t-study draw if the meta-analysis time t is not random or prespecified, but dependent on the results, as expressed in Section 3.3.4. We discuss the conventional conditional approach to meta-analysis error control in more detail in Section 3.4.1.

The other possible approach to error control is surviving over analysis times, which means that it should be valid for any upcoming analysis time t within a series. So the probability that a type-I error – ever – occurs in the accumulating series is controlled, whether the series reaches a large size or not. This is visualized in Figure 3.2 by the orange shading, and has a long run error rate that runs over series of any size, including the one-study series. This approach to error control is taken by methods for living systematic reviews such as *Trial sequential analysis* and *Sequential meta-analysis*. We discuss this approach of error control surviving over time in more detail in Section 3.4.2.

3.4.1 Error control conditioned on time

The null distributions of the common/fixed meta-analysis Z-statistic shown in Figure 3.1 are conditioned on the size of the series, which is the *time*: $T \geq t$. We can use our Accumulation Bias framework to give this distribution a general description, where we use $\phi_0(z^{(t)})$ to denote the assumed standard normal null distribution for the meta-analysis Z-score and obtain a conditional density using Bayes' rule:

$$\phi_0\left(z^{(t)} \,\middle|\, \mathscr{A}^{(t)}, T \geq t\right) = \frac{\phi_0(z^{(t)}) \cdot \mathbf{P}_0\left[\mathscr{A}^{(t)}, T \geq t \,\middle|\, z^{(t)}\right]}{\mathbf{P}_0\left[\mathscr{A}^{(t)}, T \geq t\right]} = \frac{\phi_0(z^{(t)}) \cdot \overline{A}_0\left(t \,\middle|\, z^{(t)}\right)}{\overline{A}_0(t)},$$

where we define:

$$\begin{aligned} \overline{A}_0\left(t \,\middle|\, z^{(t)}\right) &:= \mathbf{E}_0\left[A(t \mid Z_1, \ldots, Z_t) \,\middle|\, Z^{(t)} = z^{(t)}\right] \\ \overline{A}_0(t) &:= \mathbf{E}_0\left[A(t \mid Z_1, \ldots, Z_t)\right], \end{aligned} \tag{3.16}$$

with under the equal study size assumption in (3.1b)

$$Z^{(t)} = \frac{1}{\sqrt{t}} \sum_{i=1}^{t} Z_i$$

(extension to the general cases with unequal sample sizes is straightforward). For the *Gold Rush* example, $\overline{A}_0(t)$ was given by (3.10) and can be calculated if ωs are known. $\overline{A}_0(t)$ denotes the general probability of arriving at $T \geq t$ under the null hypothesis, and

so does $\bar{A}_0\left(t \,\middle|\, z^{(t)}\right)$, but with the restriction that we only take samples into account that result in meta-analysis score $z^{(t)}$. The type-I error rates for the *Gold Rush* example shown in Table 3.2 are based on a randomly chosen or prespecified t for which $\mathbf{P}[\mathscr{A}^{(t)}] = 1$, and represent the following (with ϕ_0 as above in (3.16), the standard normal density):

$$\mathbf{P}_0\left[\mathscr{E}^{(t)}_{\text{TYPE-I}} \,\middle|\, \mathscr{A}^{(t)}, T \geq t\right] = \int_{-\infty}^{-z_{\frac{\alpha}{2}}} \phi_0\left(z^{(t)} \,\middle|\, \mathscr{A}^{(t)}, T \geq t\right) dz^{(t)} + \int_{z_{\frac{\alpha}{2}}}^{\infty} \phi_0\left(z^{(t)} \,\middle|\, \mathscr{A}^{(t)}, T \geq t\right) dz^{(t)},$$

$$\phi_0\left(z_1, \ldots, z_t \,|\, \mathscr{A}\right) = \frac{\phi_0\left(z_1, \ldots, z_t\right) \cdot \bar{A}_0\left(t \,|\, z_1, \ldots, z_t\right)}{\bar{A}_0(t)}$$

where we define:

$$\begin{aligned} \bar{A}_0\left(t \,|\, z_1, \ldots, z_t\right) &:= \mathbf{E}_0\left[A(t \,|\, Z_1, \ldots, Z_t) \,|\, Z_1 = z_1, \ldots, Z_t = z_t\right] \\ \bar{A}_0(t) &:= \mathbf{E}_0\left[A(t \,|\, Z_1, \ldots, Z_t)\right] \end{aligned} \tag{3.17}$$

3.4.2 Error control surviving over time

In living systematic reviews, a meta-analysis is performed after each new study ($\mathbf{P}\left[\mathscr{A}^{(t)}\right] = 1$ for all t). The properties on error control obtained by for example *Trial Sequential Analysis* are therefore surviving over analysis times t and depend on the joint distribution on the data and the maximum study series size T. For $\mathbf{P}\left[\mathscr{A}^{(t)}\right]$ always 1, $A(t) = S(t-1)$ and this joint distribution can be presented as follows:

$$\phi_0\left(z^{(1)}, \ldots, z^{(t)}, T = t\right) = \phi_0\left(z^{(1)}, \ldots, z^{(t)}\right) \cdot \mathbf{P}_0\left[T = t \,\middle|\, z^{(1)}, \ldots, z^{(t)}\right], \tag{3.18}$$

where we define

$$\begin{aligned} \mathbf{P}_0\left[T = t \,\middle|\, z^{(1)}, \ldots, z^{(t)}\right] := \mathbf{E}_0\left[S(t-1 \,\middle|\, Z_1, \ldots, Z_{t-1}) \,\middle|\, Z^{(1)} = z^{(1)}, \ldots\right] \\ - \mathbf{E}_0\left[S(t \,\middle|\, Z_1, \ldots, Z_t) \,\middle|\, Z^{(1)} = z^{(1)}, \ldots\right], \end{aligned}$$

with under the equal study size assumption in (3.1b),

$$Z^{(t)} = \frac{1}{\sqrt{t}} \sum_{i=1}^{t} Z_i, \quad \phi_0(z^{(0)}) = 1 \text{ and } \mathbf{P}_0\left[T \geq 1 \,\middle|\, z^{(0)}, z^{(1)}\right] = 1.$$

The result $\mathbf{P}[T = t] = S(t-1) - S(t)$ is known from survival analysis and made explicit in the Appendix Section 3.E. When $S(t)$ is known for all t, it is possible to obtain error control that survives over analysis times $T = t$ with thresholds $z^{(t)}_{\frac{\alpha}{2}}$ that are functions of α, t and some $T_{\max}$ based on a maximum sample or information size. Such methods are known as *Trial sequential analysis* (Brok et al., 2008; Wetterslev et al., 2008) and *Sequential meta-analysis* (Whitehead, 2002, Ch. 12) (Whitehead, 1997; Higgins et al., 2011). If we assume a one-sided test, the approach to error control taken by these methods can be expressed

as follows:

$$\mathbf{E}_T\left[\mathbf{P}_0\left[\mathscr{E}^{(T)}_{\text{TYPE-I}}\,\middle|\,T\right]\right]=\sum_{t=1}^{T_{\max}}\int_{z^{(1)}_{\frac{\alpha}{2}}}^{\infty}\cdots\int_{z^{(t)}_{\frac{\alpha}{2}}}^{\infty}\phi_0\left(z^{(1)},\ldots,z^{(t)},T=t\right)dz^{(1)}\ldots dz^{(t)}=\alpha, \tag{3.19}$$

with ϕ_0 as above (3.18) and $T=t$ only in the case $\lambda(t)=\mathbf{1}_{Z^{(t)}\geq z^{(t)}_{\frac{\alpha}{2}}}=1$.

The change in notation from $T \geq t$ to $T = t$ already hints at the limitations of this approach: the series size needs to be completely determined by the thresholds specified in the hazard function and nothing else. We discuss this limitation in more detail in the next section.

3.4.3 Unknown and unreliable analysis time probabilities

To obtain thresholds to test $z^{(t)}$ under Accumulation Bias, we need to know the probability $A(t)$ (or only $S(t)$) for meta-analysis time t. However, any of the scenarios described in Section 3.3.3 and Section 3.3.4 can be involved, and some can be influencing $z^{(t)}$ simultaneously. Also, ethical imperatives might balance the bias, as illustrated by the following quote:

> A negative result will dampen enthusiasm and turn the attention of investigators to other possible protocols. A positive result will excite interest but may provide an ethical veto on further randomization.
>
> –Armitage (1984) as cited by Ellis and Stewart (2009)

We do not believe that the corrected thresholds $z^{(t)}_{\frac{\alpha}{2}}$ from sequential methods like *Trial Sequential Analysis* can account for all Accumulation Bias, since they require very strict conformation to the stopping rule based on synthesized studies $z^{(t)}$ and some have already argued that meta-analysts do not have such control over new studies (Chalmers and Lau, 1993). *Sequential meta-analysis* was proposed for prospective meta-analyses (Whitehead, 1997; Higgins et al., 2011) and never intended for settings with retrospective dependencies. Stopping rules based solely on meta-analysis ignore dependencies that might already have arisen at the individual study level (such as in the *Gold Rush* example) and that meta-analyses might in practice not be performed continuously (so $\mathbf{P}[\mathscr{A}^{(t)}] \neq 1$ for some t). When meta-analyses are not performed continuously, as discussed in Section 3.3.4, the specification of which series are included in the long run error control is missing (imagine for example that some of the columns 1, 2, 3 and 5 of meta-analyses in Figure 3.2 be excluded in the long run error control because the individual study results were such that nobody will ever bother to perform a meta-analysis).

It might be very inefficient to try to avoid Accumulation Bias. As stated in the introduction, avoiding it would mean that results from earlier studies should be unknown when planning new studies as well as when planning meta-analyses (that is, the decision to do a meta-analysis after t studies should not depend on the outcome of these studies). Achieving this might be impossible, since research is very often somehow inspired by other

findings. Also, such approach cannot be reconciled with the *Evidence-Based Research* initiative to reduce waste (Lund et al., 2016; Chalmers and Glasziou, 2009; Chalmers et al., 2014).

We conclude that the Accumulation Bias process specifying $A(t)$ can never be fully known and that avoiding an Accumulation Bias process will introduce more research waste. So we need a testing method that is valid regardless of the exact Accumulation Bias process. We will introduce such a method in Section 3.6, but first exhibit some evidence that, even though the recommendations from *Evidence-Based Research* still need renewed attention, Accumulation bias might already be at play.

3.5 Intermezzo: evidence for the existence of Accumulation Bias

3.5.1 Agreement with empirical findings

Accumulation Bias arises due to dependencies in how a study series comes about (Section 3.3.3), and in the timing of the meta-analysis (Section 3.3.4). We first discuss some indications of the former and then illustrate how these can be reinforced by some approaches to the latter.

If citations of previous results are a real indication of why a replication study is performed, than many such dependencies have been demonstrated in the literature on *reference/citation bias* (Gøtzsche, 1987; Egger and Smith, 1998). Citation or reference bias indicates that initial satisfactory results are more often cited than unsatisfactory results, thus some sort of *Gold Rush* occurs. Studies into citations indicate that early small trials are much more often cited than later large trials (e.g. Fergusson et al. (2005); Robinson and Goodman (2011)), which might limit the *Gold Rush* to the early studies in a series, such as indicated by Whitehead (2002) (Equation 3.3.3). Many studies have found that early studies are unreliable predictors of later replications in a study series (Roberts and Ker, 2015; Chalmers and Glasziou, 2016) (and see references 6-34 in Ioannidis (2008) and references 33-49 in Pereira and Ioannidis (2011)), which is also an indication of early study Accumulation Bias.

Other empirical findings suggest that Accumulation Bias might occur throughout a series, but to a lesser extent in later studies. Gehr et al. (2006), for example, report effect sizes that decrease over time, but in which study size did not play a significant role. What has been recognized as *regression to the truth* in heart failure studies, might also be characterized as Accumulation Bias (Krum and Tonkin, 2003). But these effects will be difficult to limit to only a few early studies, so excluding a certain number from meta-analysis, as proposed in Whitehead (2002, p. 197) (Equation 3.3.3). It is difficult to find the threshold for where the early biased studies end and the unbiased ones begin and excluding studies is a very crude measure.

The Proteus effect (Pfeiffer et al., 2011; Ioannidis and Trikalinos, 2005; Ioannidis, 2005a) describes how early replications can be biased against initial findings. If early contradicting findings spur a large series of studies into a phenomenon, it introduces a more complex pattern of Accumulation Bias that does not have a straightforward dominating direction.

The same holds for the *Value of Information* approach, to decide on replication studies (Claxton and Sculpher, 2006; Claxton et al., 2002).

There is quite some literature with suggestions on when a meta-analysis should be updated. One general recommendation is to do so when studies can be added that will have a large effect on the meta-analysis (Moher and Tsertsvadze, 2006; Moher et al., 2007b, 2008). If such recommendations reflect an overall tendency in timing of meta-analysis, Accumulation Bias might be re-enforced by the timing of the meta-analysis: initial misleading studies might have spurred a study series, and might also indirectly encourage a meta-analysis after later studies report deviating results.

3.5.2 Agreement with intuitions about priors

The famous paper "Why Most Published Research Findings are False" (Ioannidis, 2005b) introduced the concept of field specific prior odds to a large audience. The prior odds were presented as the "Ratio of True to Not-True Relationships (R)", which has the same meaning as the ratio of pilot studies from the alternative and null distribution, which we denote by $\mathbf{P}[H_1]/\mathbf{P}[H_0]$ as the prior odds. Ioannidis (2005b) combines this ratio with the average power and type-I error of tests in a research field to obtain a field-specific estimate of the Positive Predictive Value (*PPV*) of a significant result. For this, the prior odds of various research fields and publication types are given with two that are of interest to Accumulation Bias: "Adequately powered RCT with little bias" and "Confirmatory meta-analysis of good-quality RCTs". For the first of these an R of 1:1 is provided and for the second an R of 2:1. So a distinction is made between topics worthy of only one individual study and those that evoke a series of studies eligible for meta-analysis.

How would the researchers involved in replicating RCTs know that their topic is worthy of a series of studies in comparison to just one? The difference between prior odds of the two indicates that this is no random decision. The only available source of information would be previous study results, hence introducing dependence between study series size and study results: Accumulation Bias. So the prior odds R specified by Ioannidis (2005b) is actually $\frac{\mathbf{P}[H_1]\cdot\overline{A}_1(t)}{\mathbf{P}[H_0]\cdot\overline{A}_0(t)}$, with $\overline{A}_1(1)=1$ and $\overline{A}_0(1)=1$ for primary studies.

3.6 Likelihood ratios' independence from meta-analysis time

In Section 3.4.3 we argued that any approach to model the analysis time probabilities $A(t)$ is unreliable: in realistic and practically relevant scenarios, the ingredients required to calculate $A(t)$ will be unknown. Therefore, we need to define test statistics that are independent from how a series size or meta-analysis comes about. A possible form of such a test statistic is the likelihood ratio, which we discuss from the two approaches to error control: in the next Section 3.6.1 from the perspective of error control conditioned on time, and in Section 3.6.2 from the perspective of error control surviving over time.

Our proposed use of the likelihood ratio is based on the following extraordinary property[2],

[2]This property is related to the well-known fact that the Bayesian posterior based on data, when the priors are determined independently of the sample size, takes on the same value irrespective of the stopping rule that

already recognized by Berger and Berry (1988) and shown in (3.20): The likelihood ratio is a test statistic that depends on the specification of some alternative distribution ϕ_1 (a normal distribution with variance 1 just as ϕ_0) but with arbitrary mean μ other than 0. Given the data, any data sampled from an alternative distribution will have the same analysis time probabilities as data sampled from the null distribution, since analysis time probabilities are independent from the data-generating hypothesis (Section 3.3.2). When a likelihood ratio statistic is obtained for known data, the analysis time probability is a constant factor that is the same in the numerator and denominator of the likelihood ratio and therefore drops out of the equation:

$$\begin{aligned}\mathbf{LR}^{(t)}\left(z_1,\ldots,z_t,\mathscr{A}^{(t)},T\geq t\right) :=& \frac{\phi_1\left(z_1,\ldots,z_t\right)\cdot \mathbf{P}_1(\mathscr{A}^{(t)},T\geq t\mid z_1,\ldots,z_t)}{\phi_0\left(z_1,\ldots,z_t\right)\cdot \mathbf{P}_0(\mathscr{A}^{(t)},T\geq t\mid z_1,\ldots,z_t)}\\ =& \frac{\phi_1\left(z_1,\ldots,z_t\right)\cdot A(t\mid z_1,\ldots,z_t)}{\phi_0\left(z_1,\ldots,z_t\right)\cdot A(t\mid z_1,\ldots,z_t)}\\ =& \frac{\phi_1\left(z_1,\ldots,z_t\right)}{\phi_0\left(z_1,\ldots,z_t\right)}\\ =& \mathbf{LR}\left(z_1,\ldots,z_t\right).\end{aligned} \tag{3.20}$$

Here we used the standard definition of likelihood ratio for the case that the likelihood jointly involves continuous-valued data and discrete events, and we critically used the fact that the probability of $\mathscr{A}^{(t)}, T\geq t$ does not depend on whether the null or the alternative distribution generated the data.

In the following two sections we discuss two means of using likelihood-ratio based tests that yield results that are valid irrespective of accumulation bias.[3]

3.6.1 Likelihood ratio's error control conditioned on time

A large study series has an extremely low probability of occurring under the null hypothesis in the *Gold Rush* scenario, and under any other similar Accumulation Bias setting. The probability of reaching a certain study series size t is much larger under any alternative hypothesis when the power of the test for that alternative hypothesis $(1-\beta)$ is larger than the type-I error α. Due to this fact, it is possible to control an error rate if we assume that a certain fraction of pilot studies (or topics, see Figure 3.2) $\mathbf{P}[H_1]$ are sampled from the alternative distribution and a proportion $\mathbf{P}[H_0]$ of pilot studies from the null. This way, we are able to control the fraction of true rejections $1-\mathbf{P}_1\left[\mathscr{E}^{(t)}_{\text{TYPE-II}}\,\middle|\,\mathscr{A}^{(t)},T\geq t\right]$ (complement of type-II errors) to false rejections $\mathbf{P}_0\left[\mathscr{E}^{(t)}_{\text{TYPE-I}}\,\middle|\,\mathscr{A}^{(t)},T\geq t\right]$.

gave rise to the observations (Hendriksen et al., 2020)

[3] To avoid any confusion, let us highlight that our likelihood-ratio based tests are *never* equivalent to p-value based tests. While some p-value based tests (such as the Neyman-Pearson most powerful test) can be written as likelihood ratio tests, these are invariably of the form 'reject at significance level α if $\mathbf{LR}(z_1,\ldots,z_t)\geq\gamma$ where γ is chosen such that $\mathbf{P}_0(\phi_1(z_1,\ldots,z_t)/\phi_0(z_1,\ldots,z_t)\geq\gamma)=\alpha$. $\mathbf{P}_0(\phi_1(z_1,\ldots,z_t)/\phi_0(z_1,\ldots,z_t)\geq\gamma)=\alpha$. In contrast, we choose γ in a way that does not depend on knowledge of the tail area under $\mathbf{P}_0$ (e.g. in Section 3.6.2 we take $\gamma=1/\alpha$, and there the equality above is a (strict) inequality).

We can achieve such error control conditioned on time – e.g. error control taking into account only t-study meta-analyses – if we define thresholds based on the *Bayes posterior odds*, which, by Bayes' theorem, are given by $O_{\text{post}}(z_1,\ldots,z_t) = \mathbf{LR}(z_1,\ldots,z_t)\cdot\frac{\mathbf{P}[H_1]}{\mathbf{P}[H_0]}$. Remarkably, these are not affected by the mechanism underlying the decisions to continue studies or perform meta-analyses:

$$\begin{aligned} O_{\text{post}}\left(z_1,\ldots,z_t \,\middle|\, \mathscr{A}^{(t)}, T \geq t\right) :=& \frac{\mathbf{P}\left[H_1 \,\middle|\, z_1,\ldots,z_t,\mathscr{A}^{(t)}, T\geq t\right]}{\mathbf{P}\left[H_0 \,|\, z_1,\ldots,z_t,\mathscr{A}^{(t)}, T\geq t\right]} \\ =& \frac{\phi_1\left(z_1,\ldots,z_t,\mathscr{A}^{(t)}, T\geq t\right)\cdot\mathbf{P}[H_1]}{\phi_0\left(z_1,\ldots,z_t,\mathscr{A}^{(t)}, T\geq t\right)\cdot\mathbf{P}[H_0]} \\ =& \mathbf{LR}^{(t)}\left(z_1,\ldots,z_t,\mathscr{A}^{(t)}, T\geq t\right)\cdot\frac{\mathbf{P}[H_1]}{\mathbf{P}[H_0]} \\ =& \mathbf{LR}(z_1,\ldots,z_t)\cdot\frac{\mathbf{P}[H_1]}{\mathbf{P}[H_0]} = O_{\text{post}}(z_1,\ldots,z_t). \end{aligned} \tag{3.21}$$

To introduce these ideas in a simple setting, we assume here that both hypothesis consist of a single normal distribution with variance 1: $H_0 = \phi_0$ and $H_1 = \phi_1$. We can set a threshold r based on the rate of true to false rejections, so $r = 16$ would mean that we try to achieve 16 times more true rejections than false rejections $r = \frac{1-\beta}{\alpha}$, which is the usual goal of a primary analysis with intended power $1-\beta = 0.8$ and type-I error rate $\alpha = 0.05$. To obtain error control, we need to specify this r and use it to threshold the posterior odds (3.21). We define R to be the region of the sample space and $\mathscr{R}$ the event for which $O_{\text{post}}(z_1,\ldots,z_t)\geq r$, i.e. the event that we reject, and obtain the following:

$$\begin{aligned} \frac{1-\mathbf{P}_1\left[\mathscr{E}^{(t)}_{\text{TYPE-II}} \,\middle|\, \mathscr{A}^{(t)}, T\geq t\right]}{\mathbf{P}_0\left[\mathscr{E}^{(t)}_{\text{TYPE-I}} \,\middle|\, \mathscr{A}^{(t)}, T\geq t\right]}\cdot\frac{\mathbf{P}[H_1]}{\mathbf{P}[H_0]} =& \frac{\mathbf{P}_1\left[O_{\text{post}}\left(Z_1,\ldots,Z_t\middle|\mathscr{A}^{(t)}, T\geq t\right)\geq r\right]}{\mathbf{P}_0\left[O_{\text{post}}\left(Z_1,\ldots,Z_t\middle|\mathscr{A}^{(t)}, T\geq t\right)\geq r\right]}\cdot\frac{\mathbf{P}[H_1]}{\mathbf{P}[H_0]} \\ =& \frac{\mathbf{P}_1\left[O_{\text{post}}(Z_1,\ldots,Z_t)\geq r\right]}{\mathbf{P}_0\left[O_{\text{post}}(Z_1,\ldots,Z_t)\geq r\right]}\cdot\frac{\mathbf{P}[H_1]}{\mathbf{P}[H_0]} \\ =& \frac{\mathbf{P}_1[\mathscr{R}]}{\mathbf{P}_0[\mathscr{R}]}\cdot\frac{\mathbf{P}[H_1]}{\mathbf{P}[H_0]} \geq \frac{r\cdot\frac{\mathbf{P}[H_0]}{\mathbf{P}[H_1]}\cdot\mathbf{P}_0[\mathscr{R}]}{\mathbf{P}_0[\mathscr{R}]}\cdot\frac{\mathbf{P}[H_1]}{\mathbf{P}[H_0]} = r \end{aligned} \tag{3.22}$$

where the inequality follows since if $O_{\text{post}}(z_1,\ldots,z_t)\geq r$:

$$\frac{\phi_1(z_1,\ldots,z_t)}{\phi_0(z_1,\ldots,z_t)}\cdot\frac{\mathbf{P}[H_1]}{\mathbf{P}[H_0]}\geq r \qquad \text{then} \qquad \frac{\phi_1(z_1,\ldots,z_t)}{\phi_0(z_1,\ldots,z_t)}\geq r\cdot\frac{\mathbf{P}[H_0]}{\mathbf{P}[H_1]} \tag{3.23}$$

$$\text{and} \qquad \mathbf{P}_1[\mathscr{R}] = \int_R \phi_1(z_1,\ldots,z_t) \geq \int_R r\cdot\frac{\mathbf{P}[H_0]}{\mathbf{P}[H_1]}\cdot\phi_0(z_1,\ldots,z_t) = r\cdot\frac{\mathbf{P}[H_0]}{\mathbf{P}[H_1]}\cdot\mathbf{P}_0[\mathscr{R}].$$

So by specifying $\frac{\mathbf{P}[H_1]}{\mathbf{P}[H_0]}$ and an intended rate of true to false rejections r, we can calculate the posterior odds based on the likelihood ratio, compare it to the threshold based on

r and control fraction r of type-I errors under the null hypothesis. Note that any $\mathscr{A}^{(t)}$ is allowed, also multiple testing in a series or selection for the most promising meta-analysis timing. Setting a threshold to the Bayes posterior odds as described above, achieves conditional error control under any form of Accumulation Bias.

3.6.2 Likelihood ratio's error control surviving over time

A likelihood ratio itself can be used as a test statistic to obtain a procedure that controls $\mathbf{P}_0[\mathscr{E}_{\text{TYPE-I}}]$ surviving over analysis times t, as in Section 3.4.2. Suppose we simply reject if the likelihood ratio in favor of the alternative is larger than $1/\alpha$, ignoring any knowledge we might have about the accumulation bias process and the prior odds. We then find:

$$\begin{aligned}\mathbf{P}_0\Big[\text{there exists } t \leq T \text{ with } \mathscr{E}^{(t)}_{\text{TYPE-I}} \text{ and } \mathscr{A}^{(t)}\Big] &= \mathbf{P}_0\Big[\exists t \leq T:\ \mathscr{E}^{(t)}_{\text{TYPE-I}}; \mathscr{A}^{(t)}\Big] \\ &= \mathbf{P}_0\left[\exists t \leq T:\ \mathbf{LR}^{(t)}(Z_1, \ldots, Z_t) \geq \frac{1}{\alpha}; \mathscr{A}^{(t)}\right] \\ &\leq \mathbf{P}_0\left[\exists t > 0:\ \mathbf{LR}^{(t)}(Z_1, \ldots, Z_t) \geq \frac{1}{\alpha}\right] \leq \alpha. \end{aligned} \tag{3.24}$$

The final inequality is a classic result, proofs of which can be found in, for example, Robbins (1970) and (with substantial explanation) Hendriksen et al. (2020); see also Royall (2000).

Thus, the type-I error control survives over time in the sense that the $\mathbf{P}_0$-probability that we *ever* reject at a meta-analysis time is bounded by α. To further illustrate and interpret error control surviving over time, we define

$$\mathscr{F}^{(t)}_{\text{TYPE-I}} = \mathscr{E}^{(t)}_{\text{TYPE-I}} \cap \overline{\mathscr{E}}^{(t-1)}_{\text{TYPE-I}}, \cap \ldots \cap \overline{\mathscr{E}}^{(1)}_{\text{TYPE-I}}$$

as the event that the *first* type-I error $\mathscr{E}^{(t)}_{\text{TYPE-I}}$ in a series happens at time t (here $\overline{\mathscr{E}}^{(t')}_{\text{TYPE-I}}$ means 'no type-I error at time t''). As we show in Appendix Section 3.F, the previous inequality implies that

$$\sum_t \mathbf{P}_0\Big[\mathscr{F}^{(t)}_{\text{TYPE-I}}, \mathscr{A}^{(t)}, T \geq t\Big] \leq \alpha. \tag{3.25}$$

The change in notation from $\mathscr{E}^{(t)}_{\text{TYPE-I}}$ to $\mathscr{F}^{(t)}_{\text{TYPE-I}}$ is necessary since we want a general result for all forms of Accumulation Bias and do not want to assume that the series stops growing after the threshold is crossed (as is assumed in living systematic reviews, see Equation 3.3.3). But since it is not possible to control the amount of errors if multiple errors are made in the same series, we count only the first error in (3.25). As such, we are able to control the number of topics for which an error ever occurs in the series by comparing the likelihood ratio to the threshold $\frac{1}{\alpha}$.

It may seem surprising that it is possible to obtain error control in the sense of (3.25) for Accumulation Bias scenarios like *Gold Rush* example. After all, in this example large study series have only a large probability to occur if they contain many extreme (significant)

results. So it seems that we would inevitably hit a type-I error once we perform a meta-analysis. But note that in this example, the expectation of $A(t \mid Z_1, \ldots, Z_t)$ $(\overline{A}_0(t))$ is much larger for small t – due to the $S(t)$ component – so that most meta-analyses will be of small study series, or even one-study series, with small type-I error rates. In terms of Figure 3.2, controlling error this way is possible because error control runs over all topics, regardless of the realized series size. Thus, such error control is only meaningful if the series for each topic are continuously monitored – including those consisting of only pilot studies.

3.7 The choice between error control conditioned and surviving over time

Many meta-analysts seem reluctant to apply living systematic review techniques to all meta-analyses. We believe that this reluctance can be defended based on the assumed approach to error control surviving over time. Surviving over time means that all possible analysis times are weighted and that – in the long run – a large proportion of meta-analyses will be one-, two- and three-study meta-analyses and never expand. To the occasional meta-analyst, not involved in continuously updating meta-analyses, two- or three-study meta-analyses might never occur. Also, it requires a stretch of mind to imagine one-study meta-analyses part of the long run properties of your specific 15-study meta-analysis. But it has been argued that "primary research is increasingly viewed as part of a wider sequential process" (Higgins et al., 2011, p. 918), or at least, that it should be Lund et al. (2016). Whether this approach to error control is acceptable might also be very field specific. Among medical meta-analyses in the Cochrane Database of Systematic Reviews, two- and three-study meta-analyses are common Davey et al. (2011), but in other fields meta-analyses might only be performed if many more studies are available.

If, on the other hand, we want to stick to the conventional conditional approach to meta-analysis, we need additional assumptions on the fraction $\mathbf{P}[H_1]$ of true alternative hypotheses among pilot studies to threshold the posterior odds. Assuming a base rate $\mathbf{P}[H_1]$ means that we are essentially Bayesian about the null and alternative hypothesis[4], but there is no need to be strictly Bayesian: in practice, we might play around, and try best case and worst case $\mathbf{P}[H_1]$, to see how it affects our posterior odds. The important thing for us to note within the context of this chapter is that, when concentrating on posterior odds, we can ignore all details of the Accumulation Bias process and still obtain meaningful results, in the form of error control that balances type-I and type-II errors.

Summarizing: If we prefer conditional error control, we can obtain meaningful error control despite Accumulation Bias if we use tests based on likelihood ratios, but using prior odds for the base rates (and being partially Bayesian) is then unavoidable. If we prefer not to rely on any prior odds, we can still obtain meaningful error control despite Accumulation Bias if we use tests based on likelihood ratios, but then we have to resort to error control surviving over time instead of conditional error control.

[4] We do not necessarily have to be *completely* Bayesian: even if the null and/or alternative are composite, we can define "likelihood ratios" that do not rely on prior guesses about the parameters within the models. But we do need to be partially Bayesian, in the sense that we need to specify a base rate for the null (Grünwald et al., 2019). We call this pseudo-Bayes posterior odds and explain this further in the appendices to Chapter 5.

The former, conditional approach balances type-I and type-II errors and thus takes power into account. The importance of taking power (the complement of a the type-II error rate) into account has been argued before by many (Simmonds et al., 2017). In the general approach to error control in individual studies, the expected type-I error rate is fixed by the significance level α, and the type-II error rate minimized by the experimental design and sample size. In retrospective meta-analysis, however, sample size (or study series size t) is not under the control of the meta-analyst. Also, the study series size t is only a snapshot of a possibly growing series ($T \geq t$), since more studies might be performed in the future. Therefore also estimations of meta-analysis power are snapshots at a specific meta-analysis time. Nevertheless, it is often argued that many meta-analyses are underpowered (Turner et al., 2013; Davey et al., 2011) and that this should be taken into account in evaluating significance in meta-analyses. In Trial Sequential Analysis (Wetterslev et al., 2008) for example, an alternative hypothesis is formulated to judge the fraction of a required sample size available at t studies. A later review on trial sequential analysis noted:

> statistical confidence intervals and significance tests, relating exclusively to the null hypothesis, ignore the necessity of a sufficiently large number of observations to assess realistic or minimally important intervention effects.
>
> –Wetterslev, Jakobsen & Gluud (2017, p. 12)

Testing procedures based on likelihood ratios are very well suited to take an alternative distribution with minimally important intervention effect into account. Especially when balancing type-I error and power by thresholding posterior odds. Specifying power in tests without fixed sample sizes is studied extensively in Grünwald et al. (2019) and will be the focus of future research into likelihood ratios for meta-analysis.

3.8 Why likelihood ratios work: dependencies as strategy

We calculate p-values to judge the extremeness of our results under the null hypothesis, and to control type-I errors. But the p-value method is a fairly complicated approach to that goal when it comes to meta-analysis: To obtain a valid p-value for a series of studies, the sampling distribution under the null hypothesis needs to specify exactly how the series and the meta-analysis timing came about. Only for a completely and accurately specified process can the extremeness of the data be judged and compared to a threshold based on the tail area of the sampling distribution.

Fortunately, much simpler approaches to the same goal can be found. One intuitive way is to consider a series of bets $s(Z_1), s(Z_2), \ldots, s(Z_t)$ that make a profit when observed study results are extreme. The more extreme the results, the larger the profit. The bet needs to be designed in such a way that, under the null hypothesis, no profit is to be expected. Each null result costs \$1 to play the bet, but in expectation also makes a \$1 profit:

$$\mathbf{E}_0[s(Z_t)] = \$1. \tag{3.26}$$

Suppose that you start by investing \$1 in the first bet. After each study, you either decide to do a new study, and reinvest all profit obtained so far, or to stop and cash out. If you cash out after, for example, three studies, your profit is $s(Z_1) \cdot s(Z_2) \cdot s(Z_3)$.

As long as (3.26) holds for each bet, you cannot expect to profit under the null hypothesis; no matter what the process is for deciding, based on past data, to continue to new studies or to stop. This can be mathematically proven using martingale theory, but intuitively the reason is clear: The situation is entirely analogous to that in a casino where you cannot expect to make a salary out of playing – no matter how sophisticated the strategy you use on the order of the games or when you want to play or want to go home. Thus, irrespective of the rules used for continuation and stopping, making a large profit casts doubt on the null hypothesis even without knowledge of the entire sampling distribution.

This idea of testing by betting is described in great detail by Shafer and Vovk (2019), and Shafer et al. (2011) show that a likelihood ratio is a beautiful way to specify such bets. Briefly, if we set $s(Z_t) = \phi_1(Z_t)/\phi_0(Z_t)$, then (3.26) obviously holds:

$$\mathbf{E}_0\left[\frac{\phi_1(Z_t)}{\phi_0(Z_t)}\right] = \int_z \phi_0(z)\frac{\phi_1(z)}{\phi_0(z)}dz = \int_z \phi_1(z)dz = 1. \tag{3.27}$$

Under this definition, $s(z_1) \cdot \ldots \cdot s(z_t)$ has two interpretations: First, it is the joint likelihood ratio for the first t studies. Second, it is the amount of profit made by sequentially reinvesting in a bet that is not expected to make a profit under the null hypothesis.

So we can think of the meta-analyst acting at time t as earning the profit specified by the likelihood ratio of the data until the t-th study, and using that information to advise on reinvestment in future studies. This procedure will not lead to bankruptcy if the null hypothesis is true, and will therefore allow you to keep reinvesting. If the null hypothesis is not true, the better the focus of the bets – determined by how close the alternative distribution in the likelihood ratio is to the data-generating distribution – the larger the expected profit. The crucial point is that every strategy is allowed, so also the ineffective ones that produce research waste: also not taking into account earlier studies is a strategy.

This interpretation – likelihood ratios as betting strategies – explains how dependencies in the series relate to the test statistic. Any Accumulation Bias process can be considered a strategy to reinvest profit made so far, by deciding on new studies ($S(t)$), or cashing out the current profit (equivalent to performing a meta-analysis at time t and advising against further studies: $\mathscr{A}^{(t)}, T = t$). This is the intuition behind the proof of results like (3.24) and (3.25) – bounds on type-I error probability in meta-analysis – that can be derived without knowledge of the Accumulation Bias process. These bounds simply express, that under the null, a large profit is unlikely no matter what the Accumulation Bias is.

> it is always legitimate to continue betting, and this makes each individual study a more informative element of a research program or a meta-analysis
>
> –Shafer (2019, p. 2)

In contrast to an all-or-nothing test for one study, inspecting the betting profit of a study (calculating the likelihood ratio) is a way to test the data without losing the ability to build on it in future studies. The likelihood ratio has the ability to maximize the rate of growth of the evidence (the betting profit or likelihood ratio) among all studies in a series, instead of the power of a single p-value test on a prespecified series size or stopping rule Shafer

(2021). It allows for promising but inconclusive initial studies and small study series to be revisited in the light of new studies, but also to keep track of the combined evidence at any time.

In this sense, the use of likelihood ratios in meta-analysis is a statistical implementation of the goals of the *Evidence Based Research Network* (Lund et al., 2016). Choosing your bets wisely, by informing new studies by previous results is just another betting strategy. You optimize what studies to perform, and how to design and analyze them. Implementing this rationale in the statistics allows to maximize the efficiency of future research and reduce research waste (Chalmers and Glasziou, 2009).

3.8.1 Expanding likelihood ratios to *Safe Tests*

When the null hypothesis is simple, it can be shown that either using bets that satisfy (3.26) under the null or using likelihood ratios or using Bayes factors is equivalent, and the gambling approach can be viewed as a form of Bayesian inference. But for composite null (as in the t-test scenario, with unknown variance σ^2), the situation is trickier: bets that satisfy (3.26) under all distributions in the null hypotheses can still be constructed, but their relation to likelihood ratios is more complicated. The paper *Safe Testing* Grünwald et al. (2019) investigates this setting in great detail and shows that 'error control surviving over time' (Section 3.6.2) can still be obtained for general composite null.

3.9 Discussion

We need to consider *time* – study chronology and analysis timing – in meta-analysis. We need it because estimates are biased by Accumulation Bias when they assume that a t-study series is a random sample from all possible t-study series, while in fact dependencies arise in accumulating science. We also need *time* because sampling distributions are greatly affected by it, and the (p-value) tail area approach to testing is very sensitive to the shape of the sampling distribution. And we need to consider *time* because it allows for new approaches to error control that recognize the accumulating nature of scientific studies. Doing so also illustrates that available meta-analysis methods – general meta-analysis and methods for living systematic reviews – target two very different approaches to type-I error control.

We believe that the exact scientific process that determines meta-analysis time can never be fully known, and that approaches to error control need to be trustworthy regardless of it. A likelihood ratio approach to testing solves this problem and has even more appealing properties. Firstly, it agrees with a form of the stopping rule principle (Berger and Berry, 1988). Secondly, it agrees with the *Prequential principle* Dawid (1984). Thirdly, it allows for a betting interpretation Shafer and Vovk (2019); Shafer (2021): reinvesting profits from one study into the next and cashing out at any time.

But this approach still leaves us with a choice: either assume a prior probability $\mathbf{P}[H_1]$ and separate meta-analyses of various sizes from each other and individual studies, or control the type-I error rate over all analysis times t and include individual studies in the meta-analysis world. The first approach is more of a reflection of the current reality in meta-

analysis, while the second can be aligned with the goals from the *Evidence-Based Research Network* (Lund et al., 2016) and *living systematic reviews* (Simmonds et al., 2017).

Accumulation Bias itself might not need to be corrected at all, which is why we want to close this chapter with the following quote:

> the intuitive notion that bias is something bad which must be corrected for, does not even fit well within the frequentist framework. [...] one could not state "use estimate $\overline{X}$ for a fixed sample size experiment, but use $\overline{X} - c(\overline{X})$ (correcting for bias) for a sequential experiment," and retain frequentist admissibility in the "real" situation where one encounters a variety of both types of problems. The requirement of unbiasedness simply seems to have no justification.
>
> –Berger & Berry (1988, p. 67)

Code

See Appendix Section 3.G for description of simulation and visualization R code and packages used to generate the code. Code is available from Electronic Archiving System - Data Archiving and Networked Services (EASY -DANS)

EASY-DANS: Accumulation Bias in Meta-Analysis: The Need to Consider Time in Error Control. `https://doi.org/10.17026/dans-x56-qfme` Ter Schure (2019)

Appendices

3.A Common/fixed-effect meta-analysis

Here we derive (3.1a) and (3.1b), shown in (3.A.4), from the notation in Borenstein et al. (2009), specifically for the setting where means and standard deviations are reported in the study series Borenstein et al. (2009, Ch. 4). We slightly adjusted the notation by using $\overline{X}_{\text{T}}$ and $\overline{X}_{\text{P}}$ instead of $\overline{X}_1$ and $\overline{X}_2$ to indicate the treatment and placebo group estimate — to avoid confusion with the study numbering — and using $\overline{D}_i$ instead of D_i Borenstein et al. (2009, p. 22) or Y_i Borenstein et al. (2009, p. 66) as an analogy to the group study mean X_i and we denote its standard deviation as σ_{D_i}. We introduce the superscript $^{(t)}$ to emphasize a meta-analysis estimate of a series of studies 1 up to t.

Let $D_i = X_{\text{T}i} - X_{\text{P}i}$ be a random variable that denotes the difference between two observations (random or paired) from the treatment group ($X_{\text{T}i}$) and the placebo group ($X_{\text{P}i}$) in study i. Let $\hat{\sigma}_{D_i}$ be the estimate of the population standard deviation of these difference scores in study i. Following the usual assumptions of common/fixed-effect meta-analysis, no distinction is made between $\hat{\sigma}_{D_i}$ and the true σ_{D_i} Borenstein et al. (2009, p. 264) and for simplicity, we assume these standard deviations to be equal across studies:

$$\text{For all } i, j \in \{1, 2, \ldots, t\} \quad \hat{\sigma}_{D_i} = \sigma_{D_i} = \hat{\sigma}_{D_j} = \sigma_{D_j} = \sigma_D \tag{3.A.1}$$

Let $\overline{D}_i = \overline{X}_{\text{T}i} - \overline{X}_{\text{P}i}$ be the estimated treatment effect in study i, i.e. the difference between the average effect in the treatment group $\overline{X}_{\text{T}i}$ in study i and the average effect in the placebo group $\overline{X}_{\text{P}i}$ in study i. The population treatment effect is denoted by Δ, and is the difference between the population mean effects in the two groups, $\Delta = \mu_{\text{T}} - \mu_{\text{P}}$ Borenstein et al. (2009, p. 21). Let $Z_i = \frac{\overline{D}_i}{SE_{\overline{D}_i}}$ be the treatment Z-score of study i that is standardized with regard to the treatment effect standard error. (3.A.2) displays the general definition of $Z^{(t)}$, the Z-score of the combined effect estimated by a common/fixed-effect meta-analysis on studies 1 up to and including t (adapted notation from Borenstein et al. (2009, p. 66)):

$$\begin{gathered} Z^{(t)} = \frac{M^{(t)}}{SE_{M^{(t)}}} \\ M^{(t)} = \frac{\sum_{i=1}^{t} W_i \overline{D}_i}{\sum_{i=1}^{t} W_i} \quad W_i = \frac{1}{SE_{\overline{D}_i}^2} \quad SE_{M^{(t)}} = \sqrt{\frac{1}{\sum_{i=1}^{t} W_i}} \end{gathered} \tag{3.A.2}$$

Let $d_i = \frac{\overline{D}_i}{\sigma_D}$ be the Cohen's d of the treatment score in study i Borenstein et al. (2009, p. 26) — so standardized with regard to the estimated population standard deviation — and let n_i denote the sample size in the treatment and placebo arm of study i (under

the assumption that all studies have equal size study arms). Since $SE^2_{d_i} = \frac{1}{n_i}$, we let $w_i = \frac{1}{SE^2_{d_i}} = \frac{1}{\frac{1}{n_i}} = n_i$ denote the weights for d_i. Based on these weights, $M^{(t)}$ and $SE_{M^{(t)}}$ can be expressed as follows, using the fact that $\overline{D}_i = d_i\sigma_D$, $SE^2_{\overline{D}_i} = \frac{\sigma_D^2}{n_i}$, and thus $W_i = w_i\frac{1}{\sigma_D^2}$ (see also Borenstein et al. (2009, p. 82)):

$$M^{(t)} = \frac{\sum_{i=1}^{t} w_i \frac{1}{\sigma_D^2} d_i\sigma_D}{\sum_{i=1}^{t} w_i \frac{1}{\sigma_D^2}} = \frac{\sum_{i=1}^{t} w_i d_i\sigma_D}{\sum_{i=1}^{t} w_i} = \frac{\sum_{i=1}^{t} n_i d_i\sigma_D}{\sum_{i=1}^{t} n_i}$$
$$SE_{M^{(t)}} = \sqrt{\frac{1}{\sum_{i=1}^{t} w_i \frac{1}{\sigma_D^2}}} = \sqrt{\frac{\sigma_D^2}{\sum_{i=1}^{t} w_i}} = \sqrt{\frac{\sigma_D^2}{\sum_{i=1}^{t} n_i}} \tag{3.A.3}$$

With $N^{(t)} = \sum_{i=1}^{t} n_i$ and $d_i = \frac{Z_i}{\sqrt{n_i}}$, the common/fixed-effect Z-score $Z^{(t)}$ of studies i up to and including t can be derived as an average weighted by the square root of the individual study sample sizes:

$$Z^{(t)} = \frac{\frac{\sum_{i=1}^{t} n_i d_i\sigma_D}{N^{(t)}}}{\sqrt{\frac{\sigma_D^2}{N^{(t)}}}} = \frac{\sum_{i=1}^{t} n_i d_i}{\sqrt{\sum_{i=1}^{t} n_i}} = \frac{\sum_{i=1}^{t} n_i \frac{Z_i}{\sqrt{n_i}}}{\sqrt{N^{(t)}}} = \frac{\sum_{i=1}^{t} \sqrt{n_i} Z_i}{\sqrt{N^{(t)}}} = \frac{\sum_{i=1}^{t} \sqrt{n} Z_i}{\sqrt{t}\sqrt{n}} = \frac{1}{\sqrt{t}} \sum_{i=1}^{t} Z_i$$
$$\text{for } n_1 = n_2 = \ldots = n_t = n \tag{3.A.4}$$

3.B Expectation *Gold Rush* conditional pilot Z-score

Here, and in the following, we assume that there is always a first study ($\mathbf{P}[T \geq 1] = 1$).

$$\mathbf{E}_0[Z_1 \mid T \geq 2] = \frac{\mathbf{E}_0\left[Z_1 \,\middle|\, T \geq 2, Z_1 \geq z_{\frac{\alpha}{2}}\right] \cdot \mathbf{P}_0\left[T \geq 2 \,\middle|\, T \geq 1, Z_1 \geq z_{\frac{\alpha}{2}}\right] \cdot \mathbf{P}_0\left[Z_1 \geq z_{\frac{\alpha}{2}}\right]}{\mathbf{P}_0[T \geq 2]}$$
$$+ \frac{\mathbf{E}_0\left[Z_1 \,\middle|\, T \geq 2, |Z_1| < z_{\frac{\alpha}{2}}\right] \cdot \mathbf{P}_0\left[T \geq 2 \,\middle|\, T \geq 1, |Z_1| < z_{\frac{\alpha}{2}}\right] \cdot \mathbf{P}_0\left[|Z_1| < z_{\frac{\alpha}{2}}\right]}{\mathbf{P}_0[T \geq 2]}$$
$$= \frac{\mathbf{E}_0\left[Z_1 \,\middle|\, T \geq 2, Z_1 \geq z_{\frac{\alpha}{2}}\right] \cdot \omega_S^{(1)} \cdot \frac{\alpha}{2} + \mathbf{E}_0\left[Z_1 \,\middle|\, T \geq 2, |Z_1| < z_{\frac{\alpha}{2}}\right] \cdot \omega_{NS}^{(1)} \cdot (1-\alpha)}{\omega_S^{(1)} \cdot \frac{\alpha}{2} + \omega_{NS}^{(1)} \cdot (1-\alpha)} \tag{3.B.1}$$

since

$$\begin{aligned}\mathbf{P}_0[T \geq 2] &= \mathbf{P}_0\left[T \geq 2 \middle| T \geq 1, Z_1 \geq z_{\frac{\alpha}{2}}\right] \cdot \mathbf{P}_0\left[Z_1 \geq z_{\frac{\alpha}{2}}\right] \\ &\quad + \mathbf{P}_0\left[T \geq 2 \middle| T \geq 1, |Z_1| < z_{\frac{\alpha}{2}}\right] \cdot \mathbf{P}_0\left[|Z_1| < z_{\frac{\alpha}{2}}\right] \\ &= \omega_{\text{S}}^{(1)} \cdot \frac{\alpha}{2} + \omega_{\text{NS}}^{(1)} \cdot (1-\alpha)\end{aligned}$$

This expression only considers significant positive and nonsignificant results in the pilot study, since we defined in (3.2) that significant negative results have 0 probability to produce replication studies. We can replace $\mathbf{P}_0$ by $\mathbf{P}$ in the middle term of the fractions in the first two rows because *new study probabilities* are independent from the data generating distribution, as discussed in Section 3.2.3.

3.C Expectation *Gold Rush* conditional meta-analysis Z-score

For all $t \geq 2$:

$$\begin{aligned}\mathbf{E}_0\left[Z^{(t)} \,\middle|\, T \geq t\right] &= \frac{\sum_{i=1}^{t} \sqrt{n_i}\, \mathbf{E}_0[Z_i \mid T \geq t]}{\sqrt{N^{(t)}}} \\ &= \frac{\sqrt{n_1}\, \mathbf{E}_0[Z_1 \mid T \geq t] + \sum_{i=2}^{t-1} \sqrt{n_i}\, \mathbf{E}_0[Z_i \mid T \geq t] + \sqrt{n_t}\, \mathbf{E}_0[Z_t \mid T \geq t]}{\sqrt{N^{(t)}}} \\ &= \frac{\sqrt{n_1}\, \mathbf{E}_0[Z_1 \mid T \geq 2] + \sum_{i=2}^{t-1} \sqrt{n_i}\, \mathbf{E}_0[Z_i \mid T \geq i+1]}{\sqrt{N^{(t)}}}\end{aligned} \tag{3.C.1}$$

Here we use that the last study in a series under the *Gold Rush* example is unbiased and has expectation 0 under the null hypothesis. We also use that the expansion of the series beyond the next study does not influence a study's expectation in our *Gold Rush* example: for $t \geq 2$ $\mathbf{E}_0[Z_1 \mid T \geq t]$ is the same as $\mathbf{E}_0[Z_1 \mid T \geq 2]$, and for any i and $t \geq i$, $\mathbf{E}_0[Z_i \mid T \geq t]$ is the same as $\mathbf{E}_0[Z_i \mid i+1]$).

3.D Mixture variance

$$
\begin{aligned}
&\mathbf{Var}\left\{Z^{(2)} \,\middle|\, T \geq 2\right\} \\
&\quad = \frac{\alpha}{2} \cdot \omega_{\text{S}}^{(1)} \cdot \mathbf{E}_0\left[\left(Z^{(2)}\right)^2 \,\middle|\, Z_1 \geq z_{\frac{\alpha}{2}}\right] + (1-\alpha) \cdot \omega_{\text{NS}}^{(1)} \cdot \mathbf{E}_0\left[\left(Z^{(2)}\right)^2 \,\middle|\, |Z_1| < z_{\frac{\alpha}{2}}\right] \\
&\qquad - \left(\frac{\alpha}{2} \cdot \omega_{\text{S}}^{(1)} \cdot \mathbf{E}_0\left[Z^{(2)} \,\middle|\, Z_1 \geq z_{\frac{\alpha}{2}}\right] + (1-\alpha) \cdot \omega_{\text{NS}}^{(1)} \cdot \mathbf{E}_0\left[Z^{(2)} \,\middle|\, |Z_1| < z_{\frac{\alpha}{2}}\right]\right)^2 \\
&\quad = \frac{\alpha}{2} \cdot \omega_{\text{S}}^{(1)} \cdot \left(\mathbf{Var}\left\{Z^{(2)} \,\middle|\, Z_1 \geq z_{\frac{\alpha}{2}}\right\} + \mathbf{E}_0\left[Z^{(2)} \,\middle|\, Z_1 \geq z_{\frac{\alpha}{2}}\right]^2\right) \\
&\qquad + (1-\alpha) \cdot \omega_{\text{NS}}^{(1)} \cdot \left(\mathbf{Var}\left\{Z^{(2)} \,\middle|\, |Z_1| > z_{\frac{\alpha}{2}}\right\} + \mathbf{E}_0\left[Z^{(2)} \,\middle|\, Z_1 \geq z_{\frac{\alpha}{2}}\right]^2\right) \\
&\qquad - \left(\frac{\alpha}{2} \cdot \omega_{\text{S}}^{(1)} \cdot \mathbf{E}_0\left[Z^{(2)} \,\middle|\, Z_1 \geq z_{\frac{\alpha}{2}}\right] + (1-\alpha) \cdot \omega_{\text{NS}}^{(1)} \cdot \mathbf{E}_0\left[Z^{(2)} \,\middle|\, |Z_1| < z_{\frac{\alpha}{2}}\right]\right)^2 \\
&\quad = \frac{\alpha}{2} \cdot \omega_{\text{S}}^{(1)} \cdot \mathbf{Var}\left\{Z^{(2)} \,\middle|\, Z_1 \geq z_{\frac{\alpha}{2}}\right\} + (1-\alpha) \cdot \omega_{\text{NS}}^{(1)} \cdot \mathbf{Var}\left\{Z^{(2)} \,\middle|\, |Z_1| > z_{\frac{\alpha}{2}}\right\} \\
&\qquad + \frac{\alpha}{2} \cdot \omega_{\text{S}}^{(1)} \cdot \mathbf{E}_0\left[Z^{(2)} \,\middle|\, Z_1 \geq z_{\frac{\alpha}{2}}\right]^2 + (1-\alpha) \cdot \omega_{\text{NS}}^{(1)} \cdot \mathbf{E}_0\left[Z^{(2)} \,\middle|\, Z_1 \geq z_{\frac{\alpha}{2}}\right]^2 \qquad \text{(3.D.1a)} \\
&\qquad - \left(\frac{\alpha}{2} \cdot \omega_{\text{S}}^{(1)} \cdot \mathbf{E}_0\left[Z^{(2)} \,\middle|\, Z_1 \geq z_{\frac{\alpha}{2}}\right] + (1-\alpha) \cdot \omega_{\text{NS}}^{(1)} \cdot \mathbf{E}_0\left[Z^{(2)} \,\middle|\, |Z_1| < z_{\frac{\alpha}{2}}\right]\right)^2 \qquad \text{(3.D.1b)}
\end{aligned}
$$

Because squaring is a convex function, we know from Jensen's Inequality that the average squared mean (3.D.1a) is larger than the square of the average mean (3.D.1b). So the variance of the mixture is larger than the mixture of the variances.

3.E Maximum time probability

The survival function $S(t-1)$ represents the probability $\mathbf{P}[T \geq t]$. The survival function is the complement of a cumulative distribution function on maximum time or stopping times T, known in survival analysis as the *lifetime distribution function* $F(t-1)$:

$$
\begin{aligned}
S(t-1) &= 1 - F(t-1) \\
\text{with} \quad F(t-1) &= \sum_{i=0}^{t-1} \mathbf{P}[T = i]
\end{aligned}
\qquad \text{(3.E.1)}
$$

$$
\begin{aligned}
S(t-1) &= 1 - \sum_{i=0}^{t-1} \mathbf{P}[T = i] \\
S(t) &= 1 - \sum_{i=0}^{t-1} \mathbf{P}[T = i] - \mathbf{P}[T = t] \\
\text{therefore:} \quad \mathbf{P}[T = t] &= S(t-1) - S(t)
\end{aligned}
\qquad \text{(3.E.2)}
$$

3.F Error control surviving over time in terms of a sum

Let $\mathscr{F}'^{(t)}_{\text{TYPE-I}}$ be the even that both $\mathscr{F}^{(t)}$ and $T \geq t$ holds. Using in the first equality below that the events $\mathscr{F}'^{(1)}_{\text{TYPE-I}}, \mathscr{F}'^{(2)}_{\text{TYPE-I}}, \ldots$ are all mutually exclusive (so that the union bound becomes an equality), we get:

$$\begin{aligned}
\sum_t \mathbf{P}_0\left[\mathscr{F}^{(t)}_{\text{TYPE-I}}, \mathscr{A}^{(t)}, T \geq t\right] &\leq \sum_t \mathbf{P}_0\left[\mathscr{F}^{(t)}_{\text{TYPE-I}}, T \geq t\right] \\
&= \mathbf{P}_0\left[\exists t > 0 : \mathscr{F}^{(t)}_{\text{TYPE-I}}, T \geq t\right] \\
&\leq \mathbf{P}_0\left[\exists t > 0 : \mathscr{F}^{(t)}_{\text{TYPE-I}}\right] \\
&= \mathbf{P}_0\left[\exists t > 0 : \mathscr{E}^{(t)}_{\text{TYPE-I}}\right] \\
&= \mathbf{P}_0\left[\exists t > 0 : \mathbf{LR}^{(t)}(Z_1, \ldots, Z_t) \geq \frac{1}{\alpha}\right] \leq \alpha
\end{aligned}$$

where the final inequality is just the final inequality of (3.24) again. (3.25) follows.

3.G Code availability

Table 3.1, Figure 3.1 and Table 3.2 were calculated, simulated and created by R code available in the EASY-DANS repository: https://doi.org/10.17026/dans-x56-qfme (see Extended data Ter Schure (2019))

Details on the OS and version at which it were run can be found below:

- Platform: x86 64-redhat-linux-gnu
- Arch: x86 64
- OS: linux-gnu
- System: x86 64, linux-gnu
- R version: 3.5.3 (2019-03-11) Great Truth
- svn rev: 76217

The following packages were used:

- ggplot2 version 3.0.0
- graphics version 3.5.3
- grDevices version 3.5.3
- methods version 3.5.3
- stats version 3.5.3
- utils version 3.5.3

4 | Accumulation Bias: How to handle it ALL-IN

Blog post

This chapter appeared as a blog post and gives more context to the claims in Chapter 3 on accumulation bias. These claims are paradoxical, after all: How can we possibly encounter enormous bias in our meta-analysis estimates and still do valid ALL-IN inference? This blog post tries to give some intuition by introducing a very extreme and simple version of accumulation bias and showing by simulation code and plots in R what counteracts the bias in an ALL-IN analysis.[1]

An estimated 85% of global health research investment is wasted (Chalmers and Glasziou, 2009); a total of one hundred billion US dollars in the year 2009 when it was estimated. The movement to reduce this research waste recommends that previous study results be taken into account when prioritising, designing and interpreting new research (Chalmers et al., 2014; Lund et al., 2016). Yet any recommendation to increase efficiency this way requires that researchers evaluate whether the studies already available are sufficient to complete the research effort; whether a new study is necessary or wasteful. These decisions are essentially stopping rules – or rather noisy accumulation processes, when no rules are enforced – and unaccounted for in standard meta-analysis. Hence reducing waste invalidates the assumptions underlying many typical statistical procedures.

Chapter 3 details all the possible ways in which the size of a study series up for meta-analysis, or the timing of the meta-analysis, might be driven by the results within those studies. Any such dependency introduces *accumulation bias*. Unfortunately, it is often impossible to fully characterize the processes at play in retrospective meta-analysis; the bias cannot be accounted for. In this blog post we revisit an example accumulation bias pro-

[1] The introduction to this blog post is the same as in Chapter 5 as they describe the same example accumulation bias but a different approach to counteracting it.

cess, that can be one of many influencing a single meta-analysis, and use it to illustrate the following key points:

- Standard meta-analysis does not take into account that researchers decide on new studies based on other study results already available. These decisions introduce accumulation bias because the analysis assumes that the size of the study series is unrelated to the studies within; it essentially conditions on the number of studies available.
- Accumulation bias does not result from questionable research practices, such as publication bias from file-drawering a selection of results. The decision to replicate only some studies instead of all of them biases the sampling distribution of study series, but can be a very efficient approach to set priorities in research and reduce research waste.
- ALL-IN meta-analysis stands for *Anytime*, *Live* and *Leading INterim* meta-analysis. It can handle accumulation bias because it does not require a set number of studies, but performs analysis on a growing series – starting from a single study and accumulating as many studies as needed.
- ALL-IN meta-analysis also allows for continuous monitoring of the evidence as new studies arrive, even as new interim results arrive. Any decision to start, stop or expand studies is possible, while keeping valid inference and type-I error control intact. Such decisions can be strategic: increasing the value of new studies, and reducing research waste.

4.1 Our example: extreme *Gold Rush* accumulation bias

We imagine a world in which a series of studies is meta-analyzed as soon as three studies become available. Many topics deserve a first initial study, but the research field is very selective with its replications. Nevertheless, for significant results in the right direction, a replication is warranted. We call this the *Gold Rush* scenario, because after each finding of a positive significant result – the gold in science – some research group rushes into a replication, but as soon as a study disappoints, the research effort is terminated and no-one bothers to ever try again. This scenario was first proposed by Ellis and Stewart (2009) and formulated in detail and under this name in Chapter 3. Here we consider the most extreme version of the *Gold Rush* where finding a significant positive result not only makes a replication more probable, but even inevitable: the dependency of occurring replications on their predecessor's result is deterministic.

Biased *Gold Rush* sampling We denote the number of studies available on a certain topic by t. This number t can also indicate the *timing* of a meta-analysis, such that a meta-analysis can possibly occur at number of studies $t = 1, 2, 3, \ldots$ up to some maximum number of studies T. This notation follows from Chapter 3; the Technical Details at the end of this blog post make the notation involved in this blog post more explicit.

We summarize the results of individual studies into a single per-study Z-score (z_1 for the

first study, z_2 for the second, etc), such that we have the following information on a series of size t:

$$z_1, z_2, \ldots, z_t$$

We distinguish between Z-scores that are significant and in the right direction, and Z-scores that are not. A first significant positive study is indicated by $\mathbf{z_1^*}$ ($z_1 > z_\alpha$ with $z_\alpha = 1.96$ for $\alpha = 2.5\%$). A first nonsignificant or negative study is indicated by z_1^- ($z_1 \leq z_\alpha$). We use the same notation for the second and third study and limit our world to three studies (our maximum $T = 3$). After all, we meta-analyze studies on all topics and only those topics that have spurred a series of three studies. Our *Gold Rush* world consists of the following possible study series:

Gold Rush world

$$\begin{array}{rcl} z_1^- & & A(1)=0 \quad A(2)=0 \quad A(3)=0 \\ \mathbf{z_1^*}, z_2^- & & A(1)=0 \quad A(2)=0 \quad A(3)=0 \\ \mathbf{z_1^*}, \mathbf{z_2^*}, z_3^- & \rightarrow \quad z^{(3)} & A(1)=0 \quad A(2)=0 \quad A(3)=1 \\ \mathbf{z_1^*}, \mathbf{z_2^*}, \mathbf{z_3^*} & \rightarrow \quad z^{(3)} & A(1)=0 \quad A(2)=0 \quad A(3)=1 \end{array}$$

Here $A(t)$ denotes whether we accumulate *and* analyze t studies: It can be that $A(2) = 0$ and $A(3) = 0$ because we are stuck at one study, but also $A(1) = 0$ because we don't "meta-analyze" that single study. It can only be that $A(2) = 1$ if we accumulate *and* meta-analyze a two-study series and $A(3) = 1$ if we accumulate *and* meta-analyze a three-study series. In our *Gold Rush* world a very specific subset of studies accumulate into a three-study series such that they are meta-analyzed (A(3) = 1).

$z^{(3)}$ denotes the Z-score of a fixed effects meta-analysis. This meta-analysis Z-score is simply a re-normalized average and can, assuming equal (large) sample size and (known) variances in all studies, be obtained from the individual study Z-scores as follows: $z^{(3)} = \frac{1}{\sqrt{3}} \sum_{i=1}^{3} z_i$. The effects of accumulation bias are not limited to fixed-effects meta-analysis (see for example Kulinskaya et al. (2016)), but fixed-effects meta-analysis does provide us with a simple illustration for the purposes of this blog post.

We observe in our *Gold Rush* world above that the study series that are eventually meta-analyzed into a Z-score $z^{(3)}$ are a very biased subset of all possible study series. So we expect these $z^{(3)}$ scores to be biased as well. In the next section, we simulate the sampling distribution of these $z^{(3)}$ scores to illustrate this bias.

4.2 The conditional sampling distribution under extreme *Gold Rush* accumulation bias

Assume that we are in the scenario that only true null effects are studied in our *Gold Rush* world, such that any new study builds on a false-positive result. How large would the bias

be if the three-study series are simply analyzed by standard meta-analysis? We illustrate this by simulating this *Gold Rush* world using the R code below.

```
# numSim.study = number of simulated first studies
# you need 1/(0.025*0.025) = 1600 first studies for each series starting with two significant studies
# 40000 series, so 64 million studies for smooth plot (takes ~2 minutes for simulation + plotting)
numSim.study <- 64000000

Z1 <- rnorm(numSim.study)
Z2 <- rnorm(numSim.study)
Z3 <- rnorm(numSim.study)

# selection based on Gold Rush accumulation bias A(3) = 1
A3 <- which((Z1 > 1.96) & (Z2 > 1.96))

calcZmeta <- function(Zs) {
  t <- length(Zs)
  1/sqrt(t)*sum(Zs)
}

# meta Zscores for a random sample of 3-study series (you don't need all 64 million for a smooth plot)
Zmeta3 <- sapply(sample(1:numSim.study, size = 40000), function(i) calcZmeta(c(Z1[i], Z2[i], Z3[i])))

# meta Zscores for a biased sample of 3-study series, biased by GoldRush A(3) = 1
Zmeta3.A3 <- sapply(A3, function(i) calcZmeta(c(Z1[i], Z2[i], Z3[i])))

ggplot(rbind(data.frame(Zscore = Zmeta3,    GoldRush = ""),
             data.frame(Zscore = Zmeta3.A3, GoldRush = "A3"))) +
  geom_histogram(aes(x = Zscore,
                     y = ..density..,  # ..density.. normalizes by group with/without A(3) GoldRush
                     fill = GoldRush), # each with their own fill
                 alpha = 0.2, bins = 120, position = "identity") +
  scale_fill_discrete(name = "Conditioned on \n Gold Rush A(t)",
                      labels = c(bquote(z^(3)), bquote(z^(3) ~ " |  A(3) = 1")))
```

Figure 4.1. *Code to create Figure 4.2*

Theoretical sampling process: A fixed-effects meta-analysis assumes that if three studies z_1, z_2, z_3 are each sampled under the null hypothesis, each has a standard normal with mean zero and the standard normal sampling distribution also applies to the combined $z^{(3)}$ score. The R code in Figure 4.1 illustrates this sampling process: First, a large population is simulated of possible first (`Z1`), second (`Z2`) and third (`Z3`) studies from a standard normal distribution. Then in `Zmeta3` each index `i` represents a possible study series, such that `c(Z1[i], Z2[i], Z3[i])` samples an unbiased study series and `calcZmeta` calculates its fixed-effects meta-analysis Z-score $z^{(3)}$. So the large number of Z-scores in `Zmeta3` captures the unbiased sampling distribution that is assumed for fixed-effects meta-analysis $z^{(3)}$-scores.

***Gold Rush* sampling process:** In contrast, the code resulting in `A3` selects only those study series for which $A(3) = 1$ under extreme *Gold Rush* accumulation bias. So the large number of Z-scores in `Zmeta3.A3` capture a biased sampling distribution for the fixed effects meta-analysis $z^{(3)}$-scores.

Meta-analysis under *Gold Rush* accumulation bias: The final lines of code in Figure 4.1 plot two histograms of $z^{(3)}$ samples, one without and one with the *Gold Rush* $A(t)$ accumulation bias process, based on `Zmeta3` and `Zmeta3.A3` respectively. Figure 4.2 gives the result.

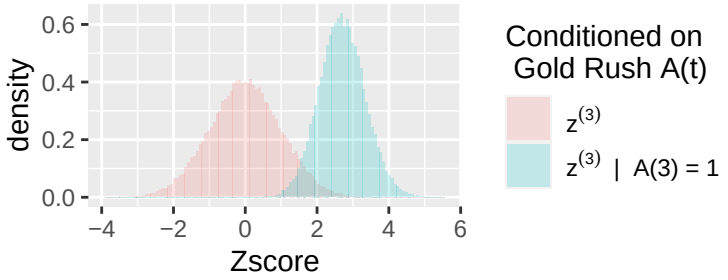

Figure 4.2. *Sampling distributions under the null hypothesis of fixed-effects meta-analysis Z-scores $Z^{(3)}$ of three studies with and without extreme* Gold Rush *accumulation bias $A(t)$, under the assumption of equal study sample size and variance.*

We observe in Figure 4.2 that the theoretical sampling process, resulting in the pink histogram, gives a distribution for the three-study meta-analysis $z^{(3)}$-scores that is centered around zero. Under the *Gold Rush* sampling process, however, our three-study $z^{(3)}$-scores do not behave like this theoretical distribution at all. The blue histogram has a smaller variance and is shifted to the right – representing the bias.

We conclude that we should not use conventional meta-analysis techniques to analyze our study series under *Gold Rush* accumulation bias: Conventional fixed-effects meta-analysis assumes that any three-study summary statistic $Z^{(3)}$ is sampled from the pink distribution in Figure 4.2 under the null hypothesis, such that the meta-analysis is significant for $Z^{(3)}$-scores larger than $z_\alpha = 1.96$ for a right-sided test with type-I error control $\alpha = 2.5\%$. Yet the actual blue sampling distribution under this accumulation bias process shows that a much larger fraction of series that accumulate three studies will have $Z^{(3)}$-scores larger than 1.96 than is assumed by the theory of random sampling. This (extremely) inflated proportion of type-I errors is 88% instead of 2.5% in our extreme *Gold Rush*, and can be obtained from our simulation by the code in Figure 4.3.

```
> typeIerror.pink <- mean(Zmeta3 > 1.96)
> typeIerror.pink
[1] 0.0250669
> typeIerror.blue <- mean(Zmeta3.A3 > 1.96)
> typeIerror.blue
[1] 0.8785025
```

Figure 4.3. *Code to calculate type-I error probability with and without extreme* Gold Rush *accumulation bias.*

4.3 Accumulation bias can be efficient

The steps in the code from Figure 4.1 that arrive at the biased distribution in Figure 4.2 illustrate that accumulation bias is in fact a selection bias. Nevertheless, accumulation bias does not result from questionable research practices, such as publication bias from file-drawering a selection of results. The selection to replicate only some studies instead

of all of them biases the sampling distribution of study series, but can be a very efficient approach to set priorities in research and reduce research waste.

By inspecting our *Gold Rush* world a bit closer, we observe that a fixed-effects meta-analysis of three studies actually *conditions* on this number of studies ($(A(t)$ needs to be $A(3)$ to be 1), and that this conditional nature is what is driving the accumulation bias; in Appendix Section 4.B we show this explicitly. In the next section we take the unconditional view.

4.4 The unconditional sampling distribution under extreme *Gold Rush* accumulation bias

We need to look at these sampling distributions from a different angle, to still achieve type-I error control. We will now introduce the unconditional sampling distribution. For this, we first adapt our *Gold Rush* accumulation bias world a bit, and not only meta-analyze three-study series but one-study "series" and two-study series as well. All possible scenarios for study series in this "all-series-size" *Gold Rush* world are illustrated below. We assume that we only meta-analyze series in a terminated state, and therefore first await a replication for significant studies before performing the meta-analysis. So a single-study "meta-analysis" can only consist of a negative or nonsignificant initial study (z_1^-); only in that case we are in a terminated state with $A(1) = 1$ and the series does not grow to two ($A(2) = 0$). In a two-study meta-analysis the series starts with a significant positive initial study and is replicated by a nonsignificant or negative one; only in that case $A(2) = 1$, and the series does not grow to three so $A(3) = 0$. And only three-study series that start with two significant positive studies are meta-analyzed in a three-study synthesis; only in that case $A(3) = 1$.

Gold Rush world; all-series-size

$$
\begin{array}{rcll}
z_1^- & \rightarrow & z^{(1)} & A(1)=1 \quad A(2)=0 \quad A(3)=0 \\
\mathbf{z_1^*}, z_2^- & \rightarrow & z^{(2)} & A(1)=0 \quad A(2)=1 \quad A(3)=0 \\
\mathbf{z_1^*}, \mathbf{z_2^*}, z_3^- & \rightarrow & z^{(3)} & A(1)=0 \quad A(2)=0 \quad A(3)=1 \\
\mathbf{z_1^*}, \mathbf{z_2^*}, \mathbf{z_3^*} & \rightarrow & z^{(3)} & A(1)=0 \quad A(2)=0 \quad A(3)=1
\end{array}
$$

The R code in Figure 4.4 calculates the fixed-effects meta-analysis $z^{(1)}$, $z^{(2)}$ and $z^{(3)}$ scores, conditional on meta-analyzing a one-study, two-study, or three-study series in this adjusted *Gold Rush* accumulation bias scenario. The histograms of these conditional $z^{(t)}$ scores are shown in Figure 4.5, including the theoretical unbiased $z^{(3)}$ histogram that was also shown in Figure 4.2 and largely overlaps with the "$A(1) = 1, A(2) = 0, A(3) = 0$"-scenario. The difference between these two sampling distributions is only visible in their right tail, with the red histogram excluding values larger than $z_\alpha = 1.96$ and redistributing their mass over other values.

Figure 4.5 clarifies that single studies are hardly biased in this extreme *Gold Rush* scenario, that the bias is problematic for two-study series and most extreme for three-study ones.

```
A1notA2 <- which(Z1 <= 1.96)
A2notA3 <- which((Z1 > 1.96) & (Z2 <= 1.96))

# meta Zscores for a biased sample  of 1-study series, biased by GoldRush A(1) = 1 and A(2) = 0
# meta Zscore of a single study is its study Zscore
Zmeta1.A1notA2 <- Z1[A1notA2]

# meta Zscores for a biased sample  of 2-study series, biased by GoldRush A(2) = 1 and A(3) = 0
Zmeta2.A2notA3 <- sapply(A2notA3, function(i) calcZmeta(c(Z1[i], Z2[i])))

ggplot(rbind(# You don't need all for a smooth plot, so sample:
             data.frame(Zscore = sample(Zmeta1.A1notA2, 40000), GoldRush = "A1notA2"),
             data.frame(Zscore = sample(Zmeta2.A2notA3, 40000), GoldRush = "A2notA3"),
             data.frame(Zscore = sample(Zmeta3.A3, 40000),      GoldRush = "A3"),
             data.frame(Zscore = sample(Zmeta3, 40000),         GoldRush = ""))) +
  geom_histogram(aes(x = Zscore, y = ..density.., # ..density.. normalizes by GoldRush group
                     fill = factor(GoldRush, levels = c("A1notA2", "A2notA3", "A3", ""))),
                 alpha = 0.2, bins = 120, position = "identity") +
  scale_fill_discrete(name = "Gold Rush A(t)",
                      labels = c(bquote(z^(1) ~ " |  A(1) = 1, A(2) = 0, A(3) = 0"),
                                 bquote(z^(2) ~ " |  A(1) = 1, A(2) = 1, A(3) = 0"),
                                 bquote(z^(3) ~ " |  A(1) = 1, A(2) = 1, A(3) = 1"),
                                 bquote(z^(3))))
```

Figure 4.4. Code to create Figure 4.5

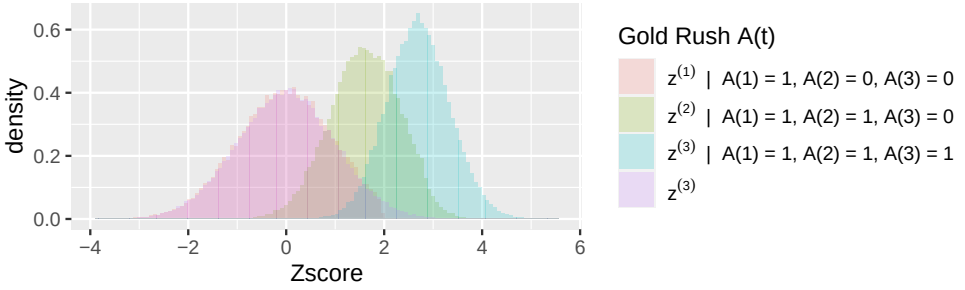

Figure 4.5. Sampling distributions under the null hypothesis of fixed-effects meta-analysis Z-scores $Z^{(t)}$ of one, two and three studies with extreme Gold Rush *accumulation bias and a three-study meta-analysis without accumulation bias, under the assumption of equal study sample size and variance.*

However, what this plot does not show us is how often we are in the one-study, two-study and three-study case.

To illustrate the relative frequencies of one-study, two-study and three-study meta-analyses, the code in Figure 4.6 samples the series in their respective numbers, instead of in equal numbers (which happens in the `sample` statements in Figure 4.4, part of creating the data frame). Plotting the total number of sampled Z-scores is dangerous for the single study $z^{(1)}$-scores, however, since there are so many of them (it can crash your R studio). So before plotting the histogram, a smaller sample is drawn that keeps the ratios between $z^{(1)}$s, $z^{(2)}$s and $z^{(3)}$s intact.

The histogram in Figure 4.7 illustrates an unconditional distribution by the raw counts of the $z^{(t)}$-scores: many result from a single study, very few from a two-study series and

```
ggplot(rbind(# You don't need all, so sample but keep ratio intact:
             data.frame(Zscore = sample(Zmeta1.A1notA2, length(A1notA2)/1000), GoldRush = "A1notA2"),
             data.frame(Zscore = sample(Zmeta2.A2notA3, length(A2notA3)/1000), GoldRush = "A2notA3"),
             data.frame(Zscore = sample(Zmeta3.A3, length(A3)/1000),           GoldRush = "A3"))) +
  geom_histogram(aes(x = Zscore, y = ..count.., # ..count.. does not normalize by GoldRush group
                     fill = factor(GoldRush, levels = c("A1notA2", "A2notA3", "A3"))),
                 alpha = 0.2, bins = 120, position = "identity") +
  scale_fill_discrete(name = "Gold Rush A(t)",
                      labels = c(bquote(z^(1) ~ " |  A(1) = 1, A(2) = 0, A(3) = 0"),
                                 bquote(z^(2) ~ " |  A(1) = 1, A(2) = 1, A(3) = 0"),
                                 bquote(z^(3) ~ " |  A(1) = 1, A(2) = 1, A(3) = 1")))
```

Figure 4.6. Code to create Figure 4.7

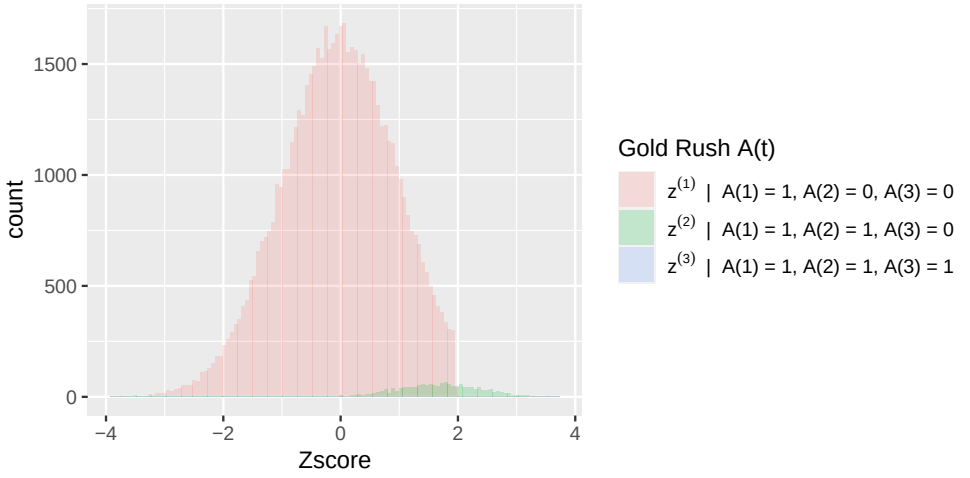

Figure 4.7. Unconditional sampling distributions under the null hypothesis of fixed-effects meta-analysis Z-scores $Z^{(t)}$ of either one, two or three studies under extreme Gold Rush *accumulation bias, under the assumption of equal study sample size and variance.*

almost none from a three-study series. In fact, this unconditional sampling distribution is hardly biased, as we will illustrate with our table further below.

We first introduce an example of an ALL-IN meta-analysis to argue that such an unconditional approach can in fact be very efficient.

4.5 ALL-IN meta-analysis

Figure 4.8 shows an example of an ALL-IN meta-analysis. Each of the red/orange/yellow lines represents a study out of the ten separate studies in as many different countries. The blue line indicates the meta-analysis synthesis of the evidence; a live account of the evidence so far in the underlying studies. In fact, *ALL-IN* meta-analysis stands for *Anytime, Live* and *Leading INterim* meta-analysis, in which the *Anytime Live* property assures valid inference under continuously monitoring and the *Leading* property allows the meta-analysis results to inform whether individual studies should be stopped or expanded.

It should be noted that such data-driven decisions would invalidate conventional meta-analysis by introducing accumulation bias.

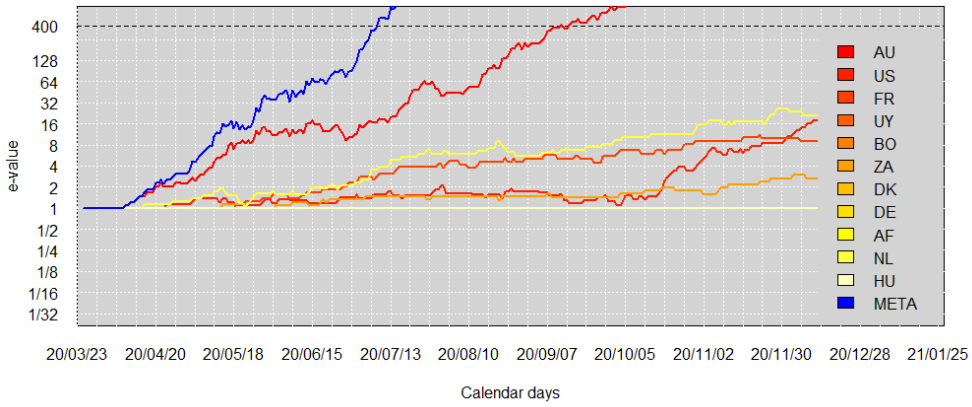

Figure 4.8. *Dashboard of an ALL-IN meta-analysis of between one and ten studies, some of which have not even started recruiting participants in the current status of this dashboard. Note that the y-axis is logarithmic.*

To interpret Figure 4.8, we observe that initially only the Australian (AU) study contributes to the meta-analysis and the blue line completely overlaps with the red one. Very quickly, the Dutch (NL) study also starts contributing and the blue meta-analysis line captures a synthesis of the evidence in two studies. Later on, also the study in the US, France (FR) and Uruguay (UY) start contributing and the meta-analysis becomes a three-study, four-study and five-study meta-analysis. How many studies contribute to the analysis, however, does not matter for its evidential value. Some studies (like the Australian one) are much larger than others, such that under a lucky scenario this study could reach the evidential threshold even before other studies start observing data. This threshold (indicated at 400) controls type-I errors at a rate of $\alpha = 1/400 = 0.0025$ (details in the final section). So in repeated sampling under the null, the combined studies will only have a probability to cross this threshold that is smaller than 0.25%. In this repeated sampling the size of the study series is essentially random: we can be lucky and observe very convincing data in the early studies, making more studies superfluous, or we can be unlucky and in need of more studies. The threshold can be reached with a single study, with a two-study meta-analysis, with a three-study,.. etc, and the repeated sampling properties, like type-I error control, hold on average over all those sampling scenarios (so unconditional on the series size).

ALL-IN meta-analysis allows for meta-analyses with Type-I error control, while completely avoiding the effects of accumulation bias and multiple testing. This is possible for two reasons: (1) we do not just perform meta-analyses on study series that have reached a

certain size, but continuously monitor study series irrespective of the current number of studies in the series; (2) we use likelihood ratios (and their cousins, e-values (Grünwald et al., 2019)) instead of raw Z-scores and p-values; we say more on likelihood ratios further below.

4.6 Accumulation bias from ALL-IN meta-analysis vs *Gold Rush*

The ALL-IN meta-analysis in Figure 4.8 illustrates an improved efficiency by not setting the number of studies in advance, but let it rely on the data and be – just like the data itself – essentially random before the start of the research effort. This introduces dependencies between study results and series size that can be expressed in similar ways as *Gold Rush* accumulation bias. Yet this field of studies might make decisions differently to our *Gold Rush*: a positive nonsignificant result might not terminate the research effort, but encourage extra studies. And instead of always encouraging extra studies, a very convincing series of significant studies might conclude the research effort. If a series of studies is dependent on any such data-driven decisions, the use of conventional statistical methods is inappropriate. These dependencies actually do not have to be extreme at all: Many fields of research might be a bit like the *Gold Rush* scenario in their response to finding significant negative results of harm. A widely known study result that indicated significant harm might make it very unlikely that the series will continue to grow. So large study series will very rarely have a completely symmetric sampling distribution, since initial studies that observe results of significant harm do not grow into large series. Hence this small aspect of accumulation bias will already invalidate conventional meta-analysis, when it assumes such symmetric distributions under the null hypothesis with equal mass on significant effects of harm and benefit.

4.7 Properties averaged over time

Accumulation bias can already result from simply excluding results of significant harm from replication. This exclusion also takes place under extreme *Gold Rush* accumulation bias, since results of significant harm as well as all nonsignificant results are not replicated. Fortunately, any such scenarios can be handled by taking an unconditional approach to meta-analysis. We will now give an intuition for why this is true in case of our extreme *Gold Rush* scenario: initial studies have bias that balances the bias in larger study series when averaged over series size and analyzed in a certain way.

Table 4.1 is inspired by Senn (2014) (different question, similar answer) and represents our extreme *Gold Rush* world of study series. It takes the same approach as Figure 4.7 and indicates the probability to meta-analyze a one-study, two-study or three-study series of each possible form under the null hypothesis. The three study series are very biased, with two or even three out of three studies showing a positive significant effect. But the $\mathbf{P}_0$ column shows that the probability of being in this scenario is very small under the null hypothesis, as was also apparent from Figure 4.7. In fact, most analysis will be of the one-study kind, that hardly have any bias, and are even slightly to the left of the theoretic standard null distribution. Exactly this phenomenon balances the biased samples of series

of larger size.

Table 4.1. Possible study series under extreme Gold Rush *accumulation bias, with their respective probabilities* $\mathbf{P}_0$ *to occur under the null hypothesis. A Z-score is marked by a * and color orange (e.g.* $\mathbf{z}_1^*$*) in case the individual study result is significant and positive (*$z_1 \geq z_\alpha$ *(one-sided test)) and by a − (e.g.* z_1^-*) otherwise. The column t indicates the number of studies and the column* $*$ *counts the number of significant studies. The fifth and sixth column multiply* $\mathbf{P}_0$ *with the* $*$ *column and t column to arrive at an expected value* $\mathbf{E}_0[*]$ *and* $\mathbf{E}_0[t]$ *respectively in the bottom row.*

t		$*$	$\mathbf{P}_0$	$* \cdot \mathbf{P}_0$	$\mathbf{t} \cdot \mathbf{P}_0$
1	z_1^-	0	$1-\alpha$	0	$1-\alpha$
2	$\mathbf{z}_1^*, z_2^-$	1	$\alpha(1-\alpha)$	$\alpha(1-\alpha)$	$2\alpha(1-\alpha)$
3	$\mathbf{z}_1^*, \mathbf{z}_2^*, z_3^-$	2	$\alpha^2(1-\alpha)$	$2\alpha^2(1-\alpha)$	$3\alpha^2(1-\alpha)$
3	$\mathbf{z}_1^*, \mathbf{z}_2^*, \mathbf{z}_3^*$	3	α^3	$3\alpha^3$	$3\alpha^3$
$\sum$			1	$\alpha+\alpha^2+\alpha^3$	$1+\alpha+\alpha^2$

The bottom row of Table 4.1 gives the expected values for the number of significant studies per series in the $* \cdot \mathbf{P}_0$ column, and the expected value for the total number of studies per series in the $t \cdot \mathbf{P}_0$ column. If we use these expressions to obtain the proportion of expected number of significant to expected total number of studies, we get the following:

$$\frac{\mathbf{E}_0[*]}{\mathbf{E}_0[t]} = \frac{\alpha+\alpha^2+\alpha^3}{1+\alpha+\alpha^2} = \frac{\alpha(1+\alpha+\alpha^2)}{1+\alpha+\alpha^2} = \alpha \tag{4.1}$$

The proportion of expected significant effects to expected series size is still α in Table 4.1 under extreme *Gold Rush* accumulation bias, as it would also be without accumulation bias.

This result is driven by the fact that there is a martingale process underlying this table. If a statistic is a martingale process and it has a certain value after t studies, the conditional expected value of the statistic after $t+1$ studies, given all the past data, is equal to the statistic after t studies. So if our proportion of significant positive studies is exactly α for the first study (t = 1), we expect to also observe a proportion α if we grow our series with an additional study (t = 1+1 = 2). The accumulation bias does not affect such statistics when averaged over time if martingales are involved (Doob's optional stopping theorem for martingales). You can verify this aspect by deleting the last row for $\mathbf{z}_1^*, \mathbf{z}_2^*, \mathbf{z}_3^*$ from our table and adding two rows for $t = 4$ in its place with $\mathbf{z}_1^*, \mathbf{z}_2^*, \mathbf{z}_3^*$ and either a fourth significant or a nonsignificant study. If you calculate the expected significant effects to expected series size, you will again arrive at α.

Martingale properties drive many approaches to sequential analysis, including the Sequential Probability Ratio Test (SPRT), group-sequential analysis and alpha spending. When applied to meta-analysis, any such inferences essentially average over series size, just like ALL-IN meta-analysis.

4.8 Multiple testing over time

Just having the expectation of some statistics not affected by stopping rules is not enough to monitor data continuously, as in ALL-IN meta-analysis. We need to account for the multiple testing as well. In that respect, the approaches to sequential analysis differ by either restricting inference to a strict stopping rule (SPRT), or setting a maximum sample size (group-sequential analysis and alpha spending).

ALL-IN meta-analysis takes an approach that is different from its predecessors and is part of an upcoming field of sequential analysis for continuous monitoring with an unlimited horizon. These approaches are called *Safe* for optional stopping and/or continuation (Grünwald et al., 2019) or any-time valid (Ramdas et al., 2020). Their methods rely on nonnegative martingales (Ramdas et al., 2020); with its most well-known and useful martingale: the likelihood ratio. For a meta-analysis Z-score, a martingale process of likelihood ratios could look as follows:

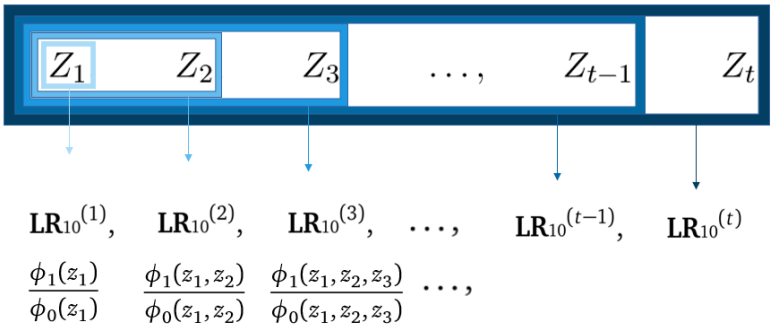

Figure 4.9. Likelihood Ratio martingale

The subscript $_{10}$ indicates that the denominator of the likelihood ratio is the likelihood of the Z-scores under the null hypothesis of mean zero, and in the numerator is some alternative mean normal likelihood. The likelihood ratio becomes smaller when the data are more likely under the null hypothesis, but the likelihood ratio can never become smaller than 0 (hence the "nonnegative" martingale). This is crucial, because a nonnegative martingale allows us to use Ville's inequality (Ville, 1939), also called the universal bound by Royall (1997). For likelihood ratios, this means that we can set a threshold that guarantees type-I error control under any accumulation bias process and at any time, as follows:

$$\mathbf{P}_0\left[\mathbf{LR}_{10}^{(t)} \geq \frac{1}{\alpha} \qquad \text{for some } t = 1, 2, \ldots\right] \leq \alpha. \tag{4.2}$$

The ALL-IN meta-analysis in Figure 4.8 in fact is based on likelihood ratios like this, and controls the type-I error by the threshold 400 at level $1/400 = 0.25\%$.

The code below illustrates that likelihood ratios can also control type-I error rates under continuous monitoring when extreme *Gold Rush* accumulation bias is at play. Within our

previous simulation, we again assume a *Gold Rush* world with only true null studies and very biased two-study and three-study series. The code in Figure 4.11 calculates likelihood ratios for the growing study series under accumulation bias. So, here we assume that a series is analyzed for each size it reaches (so after each new study), as indicated below. Figure 4.11 illustrates that still very few likelihood ratios ever grow very large.

```
numSim.study <- 10000  # we're not plotting histograms, so a smaller simulation will do

Z1 <- rnorm(numSim.study)
Z2 <- rnorm(numSim.study)
Z3 <- rnorm(numSim.study)

A1notA2 <- which(Z1 <= 1.96)
A2notA3 <- which((Z1 > 1.96) & (Z2 <= 1.96))
A3 <- which((Z1 > 1.96) & (Z2 > 1.96))

calcLR <- function(Zs) {
  prod(dnorm(Zs, mean = 1)/dnorm(Zs, mean = 0))
}

LR1.A1notA2 <- sapply(A1notA2, function(i) calcLR(Z1[i]))
LR1.A2notA3 <- sapply(A2notA3, function(i) calcLR(Z1[i]))
LR1.A3 <- sapply(A3, function(i) calcLR(Z1[i]))
LR2.A2notA3 <- sapply(A2notA3, function(i) calcLR(c(Z1[i], Z2[i])))
LR2.A3 <- sapply(A3, function(i) calcLR(c(Z1[i], Z2[i])))
LR3.A3 <- sapply(A3, function(i) calcLR(c(Z1[i], Z2[i], Z3[i])))

ggplot(rbind(data.frame(t = 0, LR = 1,           GoldRush = "A1notA2", series = A1notA2),
             data.frame(t = 0, LR = 1,           GoldRush = "A2notA3", series = A2notA3),
             data.frame(t = 0, LR = 1,           GoldRush = "A3",      series = A3),
             data.frame(t = 1, LR = LR1.A1notA2, GoldRush = "A1notA2", series = A1notA2),
             data.frame(t = 1, LR = LR1.A2notA3, GoldRush = "A2notA3", series = A2notA3),
             data.frame(t = 1, LR = LR1.A3,      GoldRush = "A3",      series = A3),
             data.frame(t = 2, LR = LR2.A2notA3, GoldRush = "A2notA3", series = A2notA3),
             data.frame(t = 2, LR = LR2.A3,      GoldRush = "A3",      series = A3),
             data.frame(t = 3, LR = LR3.A3,      GoldRush = "A3",      series = A3))) +
  geom_line(aes(x = t, y = LR, colour = GoldRush, group = series)) +
  geom_hline(yintercept = 20, linetype = "dashed") +
  scale_y_continuous(trans = 'log10') +
  scale_color_discrete(name = "Gold Rush A(t)",
                       labels = c(bquote(LR^(t) ~ " |  A(1) = 1, A(2) = 0, A(3) = 0"),
                                  bquote(LR^(t) ~ " |  A(1) = 1, A(2) = 1, A(3) = 0"),
                                  bquote(LR^(t) ~ " |  A(1) = 1, A(2) = 1, A(3) = 1")))
```

Figure 4.10. *Code to create Figure 4.11*

If we set our type-I error rate α to 5%, and compare our likelihood ratios to $1/\alpha = 20$ we observe that less than $1/20 = 5\%$ of the study series *ever* achieves a value of **LR** larger than 20 (Figure 4.12). The simulated type-I error is even much smaller than 5% since in our *Gold Rush* world series stop growing at three studies, yet this procedure controls type-I error also in the case none of these series stops growing at three studies, but all continue to grow forever.

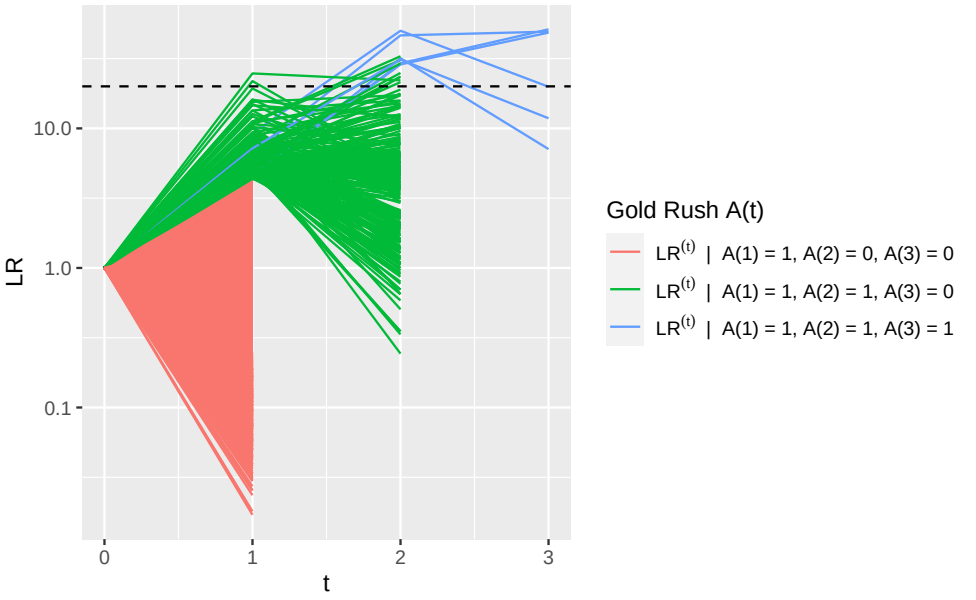

Figure 4.11. Unconditional sampling distributions under the null hypothesis of $\mathbf{LR}^{(t)}$ *of colored one, two or three studies under extreme* Gold Rush *accumulation bias. A threshold is shown at 20. Note that the y-axis is logarithmic.*

```
> typeIerrorLR <- mean(c(LR1.A1notA2,
+                        pmax(LR1.A2notA3, LR2.A2notA3),
+                        pmax(LR1.A3, LR2.A3, LR3.A3))
+                      > 20)
> typeIerrorLR
[1] 0.0023
```

Figure 4.12. Code to calculate type-I error probability for $\mathbf{LR}^{(t)}$ *averaged over series size* t *under extreme* Gold Rush *accumulation bias and continuous monitoring.*

Gold Rush world; all-series-size/continuous monitoring

$$
\begin{array}{rlll}
z_1^- & A(1)=1 & A(2)=0 & A(3)=0 \\
\mathbf{z}_1^*, z_2^- & A(1)=1 & A(2)=1 & A(3)=0 \\
\mathbf{z}_1^*, \mathbf{z}_2^*, z_3^- & A(1)=1 & A(2)=1 & A(3)=1 \\
\mathbf{z}_1^*, \mathbf{z}_2^*, \mathbf{z}_3^* & A(1)=1 & A(2)=1 & A(3)=1
\end{array}
$$

Note that continuous monitoring changes our *Gold Rush world* as indicated above. We can, however, also keep type-I error control if we do not continuously monitor, but only analyze the terminated series, such as in *Gold Rush world; all-series-size*, as shown in Figure 4.13.

The type-I error control is thus conservative, and we pay a small price in terms of power. That price is quite manageable, however, and can be tuned by setting the mean value

```
> typeIerrorLR <- mean(c(LR1.A1notA2,
+                        LR2.A2notA3,
+                        LR3.A3)
+                      > 20)
> typeIerrorLR
[1] 0.0016
```

Figure 4.13. Code to calculate type-I error probability for $\mathbf{LR}^{(t)}$ *averaged over series size* t *under extreme* Gold Rush *accumulation bias, with analysis only at the terminated series.*

of the alternative likelihood (arbitrarily set to `mean = 1` in the code for `calcLR` of Figure 4.10). More on that in Grünwald et al. (2019) and in Chapter 1.

It is this small conservatism in controlling type-I error that allows for full flexibility: There isn't a single accumulation bias process that could invalidate the inference. Any data-driven decision is allowed. And data-driven decisions can increase the value of new studies and reduce research waste.

4.9 Conclusion

In our imaginary world of extreme *Gold Rush* accumulation bias, the sampling distribution of the meta-analysis Z-score behaves very different from the sampling distribution assumed to calculate p-values and confidence intervals. A meta-analysis p-value conditions on the available sample size – on the sample size of the studies and on the number of studies available – and represents the tail area of this conditional sampling distribution under the null based on the observed Z-statistic. Analogously, a meta-analysis confidence interval provides coverage under repeated sampling from this conditional distribution. So if this sample size is driven by the data, as in any accumulation bias process, there is a mismatch between the assumed sampling distribution of the meta-analysis Z-statistic, and the actual sampling distribution.

We believe that some accumulation bias is at play in almost any retrospective meta-analysis, such that p-values and confidence intervals generally do not have their promised type-I error control and coverage. ALL-IN meta-analysis based on likelihood ratios can handle accumulation bias, even if the exact process is unknown. It also allows for continuous monitoring; multiple testing is no problem. Hence taking the ALL-IN perspective on meta-analysis will reduce research waste by allowing efficient data-driven decisions – not letting them invalidate the inference – and incorporating single studies and small study series into meta-analysis inference.

Code availability

This blogpost's R code is available on `https://osf.io/p2rtw/` (Ter Schure, 2021a).

Appendices

This blog post discusses approaches to meta-analysis that control type-I error averaged over study series size. This is called error control *surviving over time* in Chapter 3, as will become more clear in the technical details below.

Time: timing and chronology

Following notation from Chapter 3, we denote the number of studies available on a certain topic by t. This number t can also indicate the *timing* of a meta-analysis, such that a meta-analysis can possibly occur at time $t = 1, 2, 3, \ldots$ up to some maximum number of studies T. The number of studies and the timing of a meta-analysis share the notion of chronology; of past, present and future studies. At $t = 3$, we have three studies available that we can possibly meta-analyse. The fact that a third study exists can depend on the result of the first and second, but can never depend on the result of a future fourth study. Analogously, our timing of a meta-analysis after three studies can depend on the results of those three studies, but never on future meta-analyses. Note that dependencies in time are *possible*, but not necessary, to apply the notation from the accumulation bias framework (Chapter 3). Simultaneous studies can also be described, in which case their existence cannot depend on each other. A "no dependency"-relation does require the simultaneous studies to be assigned an arbitrary chronology, but their order plays no further role than to express a set of studies as a series. In the example of this blog post, however, the extreme *Gold Rush* scenario, we assume a very real chronology and deterministic dependency between all the studies in a series.

4.A Extreme *Gold Rush* expressed in accumulation bias framework

$A(t)$ denotes the probability that t studies accumulate and are analysed together in a meta-analysis. $A(t)$ has two components, the first indicates whether the topic "survives" the $(t-1)^{\text{th}}$ study, in which case the maximum number of studies T is larger than $t-1$ ($T \geq t$ or $T > t-1$ captured by the survival function $S(t-1)$), and the second indicates whether we bother to meta-analyze the series at its size t (the event $\mathscr{A}^{(t)}$). In our extreme *Gold Rush* world we assume that only three-study series are synthesized in a meta-analysis, such that $\mathbf{P}[\mathscr{A}^{(t)}]$ is only 1 for $\mathscr{A}^{(3)}$ and always 0 for $\mathscr{A}^{(2)}$ and $\mathscr{A}^{(1)}$ (we do not perform any 2-study or 1-study meta-analyses). In general, $A(t)$ depends on $S(t)$ and $\mathscr{A}^{(t)}$ as follows:

$$A(t \mid z_1 \ldots z_t) = \mathbf{P}[\mathscr{A}^{(t)} \mid T \geq t, z_1 \ldots z_t] \cdot S(t-1 \mid z_1 \ldots z_{t-1}) \tag{4.A.1}$$

In this simplified version of the *Gold Rush* scenario $S(t)$ is always either 0 or 1 if the study

results z_1 and z_2 are known:

$$S(1 \mid z_1) = \begin{cases} 1, & \text{if } \quad z_1 \text{ is of the form } \mathbf{z_1^*}. \\ 0, & \text{otherwise.} \end{cases}$$
$$S(2 \mid z_1, z_2) = \begin{cases} 1, & \text{if } \quad z_1, z_2 \text{ are of the form } \mathbf{z_1^*}, \mathbf{z_2^*}. \\ 0, & \text{otherwise.} \end{cases} \tag{4.A.2}$$

$S(t)$ is a survival probability, because a series can only grow to three studies (it has survive the second study ($S(2)$)) if it has also grown to two studies (it has survived the first study ($S(1)$). In contrast to the deterministic extreme *Gold Rush* in this paper, Chapter 3 describes a probabilistic version where the probability of a replication study is larger following a significant positive result, but not always zero following a nonsignificant one. In that case we specify *hazards* of stopping after observing a certain result, which are probabilities of stopping, given that the series accumulated so far. The survival probability is defined in terms of these hazards, following standard survival analysis notation.

In our extreme *Gold Rush* any meta-analysis has zero probability to occur except for the three-study meta-analysis ($\mathbf{P}[\mathscr{A}^{(3)}] = 1$ and for all other t $\mathbf{P}[\mathscr{A}^{(t)}] = 0$ independent of the observed results $z_1, z_2, \ldots z_t$), we find nonzero $A(t)$ only for the last two scenarios below:

$$A(2 \mid z_1^-) = \mathbf{P}[\mathscr{A}^{(2)} \mid T \geq 2] \cdot S(1 \mid z_1^-) = 0 \cdot 0 = 0 \tag{4.A.3}$$
$$A(2 \mid \mathbf{z_1^*}, z_2^-) = \mathbf{P}[\mathscr{A}^{(2)} \mid T \geq 2] \cdot S(1 \mid \mathbf{z_1^*}) = 0 \cdot 1 = 0 \tag{4.A.4}$$
$$A(2 \mid \mathbf{z_1^*}, \mathbf{z_2^*}) = \mathbf{P}[\mathscr{A}^{(2)} \mid T \geq 2] \cdot S(1 \mid \mathbf{z_1^*}) = 0 \cdot 1 = 0 \tag{4.A.5}$$
$$A(3 \mid \mathbf{z_1^*}, \mathbf{z_2^*}, z_3^- \mid T \leq 3) = \mathbf{P}[\mathscr{A}^{(3)} \mid T \geq 3] \cdot S(2 \mid \mathbf{z_1^*}, \mathbf{z_2^*}) = 1 \cdot 1 = 1 \tag{4.A.6}$$
$$A(3 \mid \mathbf{z_1^*}, \mathbf{z_2^*}, \mathbf{z_3^*}) = \mathbf{P}[\mathscr{A}^{(3)} \mid T \geq 3] \cdot S(2 \mid \mathbf{z_1^*}, \mathbf{z_2^*}) = 1 \cdot 1 = 1 \tag{4.A.7}$$

4.B Extreme *Gold Rush* conditional sampling distribution

The sampling distribution of $Z^{(t)}$ under accumulation bias is a distribution that conditions on having t studies available and analyzing them, which happens with probability $A(t)$ given the data.

Using notation from Chapter 3 we express the accumulation of t studies as $T \geq t$, indicating that once we have t studies available, our maximum amount of studies T is at least t (it is either t or larger). $\mathscr{A}^{(t)}$ indicates the event that we perform a meta-analysis of the t studies available. We denote the conditional sampling distribution of a $z^{(t)}$-score by $\phi_0(z^{(t)} \mid \mathscr{A}^{(t)}, T \geq t)$, and obtain its expression by observing that $A(t)$ is a probability of a t-study meta-analysis conditioned on the data, and we need a probability of the

data conditioned on the occurrence of a t-study meta-analysis; Bayes' rule transposes the conditional in the following expression:

$$\begin{aligned}\phi_0\left(z^{(t)} \,\middle|\, \mathscr{A}^{(t)}, T \geq t\right) &= \frac{\phi_0(z^{(t)}) \cdot \mathbf{P}_0\left[\mathscr{A}^{(t)}, T \geq t \,\middle|\, z^{(t)}\right]}{\mathbf{P}_0\left[\mathscr{A}^{(t)}, T \geq t\right]} \\ &= \frac{\phi_0(z^{(t)}) \cdot \overline{A}_0\left(t \,\middle|\, z^{(t)}\right)}{\overline{A}_0(t)},\end{aligned} \tag{4.B.1}$$

where we define:

$$\begin{aligned}\overline{A}_0\left(t \,\middle|\, z^{(t)}\right) &:= \mathbf{E}_0\left[A(t \mid Z_1, \ldots, Z_t) \,\middle|\, Z^{(t)} = z^{(t)}\right] \\ \overline{A}_0(t) &:= \mathbf{E}_0\left[A(t \mid Z_1, \ldots, Z_t)\right].\end{aligned}$$

Here, ϕ_0 is a standard normal distribution with mean 0 and variance 1. For the extreme *Gold Rush* scenario of this paper, and the sampling distribution of a three-study series illustrated in Figure 2, $\overline{A}_0(3)$ can be calculated as follows:

$$\begin{aligned}\overline{A}_0(3) &= \mathbf{E}_0[A(3 \mid Z_1, Z_2, Z_3)] \\ &= A(3 \mid \mathbf{z}_1^*, \mathbf{z}_2^*, z_3^-) \cdot \mathbf{P}_0[\mathbf{z}_1^*, \mathbf{z}_2^*, z_3^-] + A(3 \mid \mathbf{z}_1^*, \mathbf{z}_2^*, \mathbf{z}_3^*) \cdot \mathbf{P}_0[\mathbf{z}_1^*, \mathbf{z}_2^*, \mathbf{z}_3^*] \\ &= 1 \cdot \mathbf{P}_0[\mathbf{z}_1^*, \mathbf{z}_2^*, z_3^-] + 1 \cdot \mathbf{P}_0[\mathbf{z}_1^*, \mathbf{z}_2^*, \mathbf{z}_3^*] \\ &= 1 \cdot \alpha \cdot \alpha \cdot (1-\alpha) + 1 \cdot \alpha \cdot \alpha \cdot \alpha \\ &= \frac{1}{1600} \quad (\text{for } \alpha = 2.5\%)\end{aligned} \tag{4.B.2}$$

The only three-study series that have nonzero $A(t)$ are $A(3 \mid \mathbf{z}_1^*, \mathbf{z}_2^*, z_3^-)$ and $A(3 \mid \mathbf{z}_1^*, \mathbf{z}_2^*, \mathbf{z}_3^*)$, such that only these have to be enumerated in expectation $\overline{A}(3)$. $\overline{A}_0\left(3 \,\middle|\, z^{(t)}\right)$ can be obtained by considering all the possible combinations of Z_1, Z_2, Z_3 that could be summarized into a specific $z^{(t)}$ and taking into account their probabilities under the null hypothesis.

The value 1/1600 explains the statement in the beginning of the code in Figure 1 that 1600 first studies are needed for each sample of a three-study series.

A(t) behaves like a survival probability

Table 4.C.2 is an extension of the table in the blog post and shows that even though $\overline{A}_0(t)$ indicates the null hypothesis probability of accumulating t studies and meta-analyzing them, it cannot in itself tell us how often the research effort is terminated at exactly those t studies. This is caused by the fact that $A(t)$ is partly a survival probability and can be illustrated by adding a column of $\overline{A}_0(t)$ values to our table that does not add up to one.

Table 4.C.2. Possible study series under extreme Gold Rush *accumulation bias.*

	τ		$\mathbf{N}^*(\tau)$	$\overline{\mathbf{A}}_\mathbf{0}(\tau)$	$\mathbf{P}_0$	$\mathbf{N}^*(\tau)\cdot\mathbf{P}_0$	$\tau\cdot\mathbf{P}_0$
	1	z_1^-	0	1	$1-\alpha$	0	$1-\alpha$
	2	$\mathbf{z}_1^*, z_2^-$	1	α	$\alpha(1-\alpha)$	$\alpha(1-\alpha)$	$2\alpha(1-\alpha)$
	3	$\mathbf{z}_1^*, \mathbf{z}_2^*, z_3^-$	2	α^2	$\alpha^2(1-\alpha)$	$2\alpha^2(1-\alpha)$	$3\alpha^2(1-\alpha)$
$T=3$	3	$\mathbf{z}_1^*, \mathbf{z}_2^*, \mathbf{z}_3^*$	3	α^3	α^3	$3\alpha^3$	$3\alpha^3$
$\sum$					1	$\alpha+\alpha^2+\alpha^3$	$1+\alpha+\alpha^2$

4.C The martingale underlying the table

Table 4.C.2 is slightly modified in comparison to the blog post to introduce more formal notation for the *Gold Rush* stopping rule. Here we show the specific martingale underlying this table and how Doob's Optional Stopping Theorem explains the relation between the values in the bottom row of the table.

We assume that each individual study Z-score is independently sampled from a standard normal distribution with mean zero, such that the probability of obtaining a significant and positive result ($\mathbf{z}^*$ if $z \geq z_\alpha$) is α. Using $\mathbf{1}_{\mathbf{z}^*}(z_i)$ for the indicator function that indicates whether z_i is significant and positive, the martingale $\{M_1, M_2, M_3, \ldots\}$ underlying Table 4.C.2 is defined as follows:

$$M_t = \sum_{i=1}^{t} \mathbf{1}_{\mathbf{z}^*}(z_i) - t\alpha.$$

$\{M_1, M_2, M_3, \ldots\}$ is a martingale since

$$\begin{aligned} \mathbf{E}_0[M_t - M_{t-1}] &= \mathbf{E}_0\left[\sum_{i=1}^{t} \mathbf{1}_{\mathbf{z}^*}(z_i) - t\alpha - \left(\sum_{i=1}^{t-1} \mathbf{1}_{\mathbf{z}^*}(z_i) - (t-1)\alpha\right)\right] \\ &= \mathbf{E}_0[\mathbf{1}_{\mathbf{z}^*}(z_t) - \alpha] = \alpha - \alpha = 0. \end{aligned} \tag{4.D.1}$$

We denote the number of significant positive studies in a series of size t by $N^*(t)$ in Table 4.C.2 and express this number in terms of M_t:

$$N^*(t) = \sum_{i=1}^{t} \mathbf{1}_{\mathbf{z}^*}(z_i) = M_t + t\alpha.$$

The *Gold Rush* stopping rule implies that we only stop accumulating studies at series size $t = \tau$ if we find the first nonsignificant study (z_τ^-, where $\tau = \min_t\{\mathbf{1}_{\mathbf{z}^*}(z_t) = 0\}$) or if we arrive at the maximum series size $t = T$. This stopping rule forces us to stop accumulating studies at either series size τ or at size T, whichever comes first. So we stop at $\tau \wedge T$. We

express the expected number of studies that is significant and positive in terms of the expectation of the martingale under this stopping rule:

$$\mathbf{E}_0[N^*(\tau \wedge T)] = \mathbf{E}_0[M_{\tau \wedge T}] - \mathbf{E}_0[\tau \wedge T]\alpha,$$

and since $\tau \wedge T$ is always finite, by Doob's Optional Stopping theorem we have:

$$\mathbf{E}_0[M_{\tau \wedge T}] = \mathbf{E}_0[M_1] = \mathbf{E}_0\left[\sum_{i=1}^{1} \mathbf{1}_{z^*}(z_i) - 1\alpha\right] = \mathbf{E}_0[\mathbf{1}_{z^*}(z_i)] - \alpha = \alpha - \alpha = 0$$

such that

$$\mathbf{E}_0[N^*(\tau \wedge T)] = \mathbf{E}_0[\tau \wedge T]\alpha \tag{4.D.2}$$

and

$$\frac{\mathbf{E}_0[N^*(\tau \wedge T)]}{\mathbf{E}_0[\tau \wedge T]} = \alpha.$$

This is shown for the *Gold Rush* stopping rule and $T = 3$ by Table 4.C.2, but holds for any stopping rule and finite T.

5 | Accumulation Bias: How to handle it as a Bayesian

Blog post

This chapter appeared as a blog post and gives more context to the claims in Chapter 3 on accumulation bias. These claims are paradoxical, after all: how can we possibly encounter enormous bias in our meta-analysis estimates and still do valid Bayesian inference? This blog post tries to give some intuition by introducing a very extreme and simple version of accumulation bias and showing by simulation code and plots in R what counteracts the bias in a Bayesian analysis.[1]

An estimated 85% of global health research investment is wasted (Chalmers and Glasziou, 2009); a total of one hundred billion US dollars in the year 2009 when it was estimated. The movement to reduce this research waste recommends that previous study results be taken into account when prioritizing, designing, and interpreting new research (Chalmers et al., 2014; Lund et al., 2016). Yet any recommendation to increase efficiency this way requires that researchers evaluate whether the studies already available are sufficient to complete the research effort; whether a new study is necessary or wasteful. These decisions are essentially stopping rules – or rather noisy accumulation processes, when no rules are enforced – and unaccounted for in standard meta-analysis. Hence reducing waste invalidates the assumptions underlying many typical statistical procedures.

Chapter 3 details all the possible ways in which the size of a study series up for meta-analysis, or the timing of the meta-analysis, might be driven by the results within those studies. Any such dependency introduces *accumulation bias*. Unfortunately, it is often impossible to fully characterize the processes at play in retrospective meta-analysis; the bias cannot be accounted for.

[1] The introduction to this blog post is the same as in Chapter 4 as they describe the same example accumulation bias but a different approach to counteracting it.

This is the second blog post about this type of bias and how to handle it. The first blog post (Chapter 4) detailed how it can be that ALL-IN meta-analysis handles accumulation bias. This second blog post deals with the Bayesian approach. We revisit the same example accumulation bias process, which can be one of many influencing a single meta-analysis, and use it to illustrate the following key points:

- Standard meta-analysis does not take into account that researchers decide on new studies based on other study results already available. These decisions introduce accumulation bias because the analysis assumes that the size of the study series is unrelated to the studies within; it essentially conditions on the number of studies available.
- A Bayesian analysis also conditions on the number of studies available, but can still handle accumulation bias well because it compares the biased sampling distribution under the null hypothesis to those under the alternative hypothesis.
- A Bayesian analysis can have error control under accumulation bias, but this crucially depends on the ratio of null and alternative hypotheses: the prior odds. No Bayesian analysis can handle accumulation bias when the prior odds cannot be specified.
- Specifying prior odds might be difficult for a meta-analysis in retrospect: if information from the study results included in the meta-analysis seeps into the prior odds, they become invalid.
- The e-values that follow from ALL-IN meta-analysis can also be combined with prior odds in a Bayesian analysis. They combine into pseudo-Bayes posterior odds that allow Bayesian error control. By using e-values rather than standard Bayes factors we can avoid specifying prior densities on the parameters within the null and the alternative; but prior odds on H_0 and H_1 are still needed and have to be trusted.
- If trustworthy prior odds can be specified, pseudo-Bayes posterior odds allow for continuous monitoring of the evidence as new studies arrive, even as new interim results arrive. Any decision to start, stop or expand studies is possible while keeping valid inference and Bayesian error control intact. Such decisions can be strategic: increasing the value of new studies, and reducing research waste.

5.1 Our example: extreme *Gold Rush* accumulation bias

We imagine a world in which a series of studies is meta-analyzed as soon as three studies become available. Many topics deserve a first initial study, but the research field is very selective with its replications. Nevertheless, for significant results in the right direction, a replication is warranted. We call this the *Gold Rush* scenario because after each finding of a positive significant result – the gold in science – some research group rushes into a replication, but as soon as a study disappoints, the research effort is terminated and no one bothers to ever try again. This scenario was first proposed by Ellis and Stewart (2009) and formulated in detail and under this name in Chapter 3. Here we consider the most extreme version of the *Gold Rush* where finding a significant positive result not

only makes replication more probable but even inevitable: the dependency of occurring replications on their predecessor's result is deterministic.

The first blog post gave a precise definition of this extreme *Gold Rush accumulation bias* and showed by simulation that the sampling distribution under the null hypothesis is affected by such a process or stopping rule. This is shown in Figure 5.1 for the fixed-effect meta-analysis $z^{(3)}$-scores for a three-study series. The theoretical sampling process, in the pink histogram, is centered around zero and the blue histogram, under accumulation bias process $A(t)$, does not behave like this theoretical distribution at all. It has a smaller variance and is shifted to the right – representing the bias. Here $A(3) = 1$ indicates that we accumulate and analyze 3 studies under the *Gold Rush* process. (For a precise definition, please refer to the blog post Accumulation Bias: How to handle it ALL-IN in Chapter 4.)

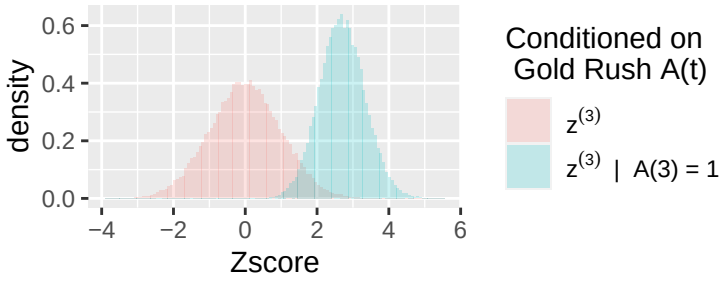

Figure 5.1. *Sampling distributions under the null hypothesis of fixed-effects meta-analysis Z-scores $Z^{(3)}$ of three studies with and without extreme* Gold Rush *accumulation bias $A(t)$, under the assumption of equal study sample size and variance.*

Bayesians claim that they can deal with any such stopping rules. So how can this be when the sampling distribution in Figure 5.1 is so much affected?

5.2 Likelihood ratios

We first turn our attention from the meta-analysis $Z^{(3)}$ statistic for three studies, to a likelihood ratio statistic $\mathbf{LR}^{(3)}$ for three studies. We summarize the results of individual studies into a single per-study Z-score (z_1 for the first study, z_2 for the second, etc), where we follow the same procedure that generated Figure 5.1, but calculate for each sample a likelihood ratio **LR** of two standard normal distributions, one with unit variance and mean 1 (ϕ_1) and one with unit variance and mean 0 (ϕ_0):

$$\mathbf{LR}^{(3)} = \frac{\phi_1(z_1, z_2, z_3)}{\phi_0(z_1, z_2, z_3)} = \prod_{i=1}^{3} \frac{\phi_1(z_i)}{\phi_0(z_i)}.$$

Assume that we are in the scenario that only true null effects are studied in our *Gold Rush* world, such that any new study builds on a false-positive result. How large would the

bias be in our likelihood ratio statistic $\mathbf{LR}^{(3)}$ if we analyze at the three-study series? We illustrate this by simulating this *Gold Rush* world using the R code below.

```
# numSim.study = number of simulated first studies
# you need 1/(0.025*0.025) = 1600 first studies for each series starting with two significant studies
# 50000 series, so 80 milion studies for smooth plot (takes ~4 minutes for simulation + plotting)
numSim.study <- 80000000

Z1 <- rnorm(numSim.study)
Z2 <- rnorm(numSim.study)
Z3 <- rnorm(numSim.study)

# selection based on Gold Rush accumulation bias A(3) = 1
A3 <- which((Z1 > 1.96) & (Z2 > 1.96))

calcLRmeta <- function(Zs) {
  prod(dnorm(Zs, mean = 1)/dnorm(Zs, mean = 0))
}

# meta LRscores for a random sample of 3-study series (you don't need all for a smooth plot)
LRmeta3 <- sapply(sample(1:numSim.study, size = 50000), function(i) calcLRmeta(c(Z1[i], Z2[i], Z3[i])))

# meta LRscores for a biased sample of 3-study series, biased by GoldRush A(3) = 1
LRmeta3.A3 <- sapply(A3, function(i) calcLRmeta(c(Z1[i], Z2[i], Z3[i])))

ggplot(rbind(data.frame(LR = LRmeta3,    GoldRush = ""),
             data.frame(LR = LRmeta3.A3, GoldRush = "A3"))) +
  geom_histogram(aes(x = LR,
                     y = ..density..,  # ..density normalizes by group with/without A(3) GoldRush
                     fill = GoldRush), # each with their own fill
                 bins = 120, position = "identity") +
  scale_x_continuous(trans = 'log10') +
  scale_fill_manual(values = hcl(15, 100, 65, alpha = c(0.8, 0.3)),
                    name = "Conditioned on \n Gold Rush A(t)",
                    labels = c(bquote(LR^(3)), bquote(LR^(3) ~ " |  A(3) = 1")))
```

Figure 5.2. Code to create Figure 5.3

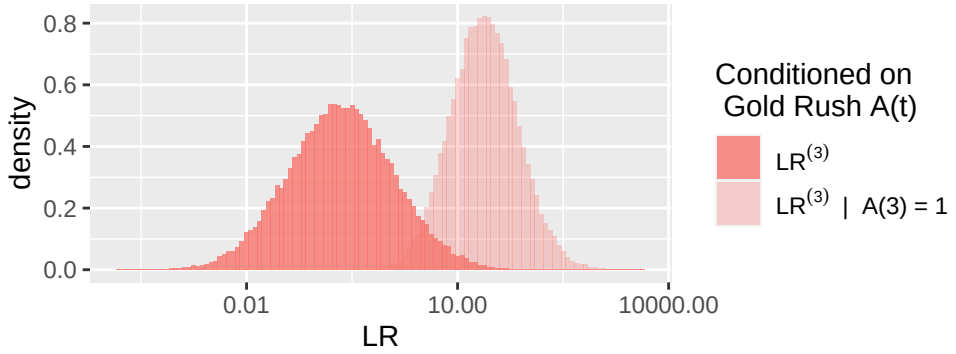

Figure 5.3. Sampling distributions under the null hypothesis of likelihood ratios $\mathbf{LR}^{(3)} = \prod_{i=1}^{3} \phi_1(Z_i)/\phi_0(Z_i)$ *of three studies with and without extreme* Gold Rush *accumulation bias* $A(t)$*. Note that the x-axis is on a log scale.*

Theoretical sampling process: A log-likelihood ratio of standard normal data has a normal sampling distribution. The R code in Figure 5.2 illustrates this sampling process:

First, a large population is simulated of possible first (`Z1`), second (`Z2`) and third (`Z3`) studies from a standard normal distribution. In the line of code for `LRmeta3`, each index `i` represents a possible study series, such that `c(Z1[i], Z2[i], Z3[i])` samples an unbiased study series and `calcLRmeta` calculates its likelihood ratio $\mathbf{LR}^{(3)}$. So the large number of Z-scores in `LRmeta3` captures the unbiased sampling distribution of the likelihood ratios.

***Gold Rush* sampling process:** In contrast, the code resulting in `A3` selects only those study series for which $A(3) = 1$ under extreme *Gold Rush* accumulation bias. So the large number of **LR**-scores in `LRmeta3.A3` capture a biased sampling distribution for $\mathbf{LR}^{(3)} \mid A(3) = 1$.

Likelihood ratios under *Gold Rush* accumulation bias: The final lines of code in Figure 5.2 plot two histograms of $\mathbf{LR}^{(3)}$ samples, one without and one with the *Gold Rush* $A(t)$ accumulation bias process, based on `LRmeta3` and `LRmeta3.A3` respectively. Each is given on the log-scale such that their normal sampling distributions become apparent. Figure 5.3 gives the result.

Here the likelihood ratio is just another statistic, with a sampling distribution that is affected by the *Gold Rush* decision making. The sampling distributions for $\mathbf{LR}^{(3)}$ on a log-scale (so $\log \mathbf{LR}^{(3)}$) in Figure 5.3 look very similar to those for $Z^{(3)}$ in Figure 5.1.

5.3 Two simple hypotheses

A Bayesian does not only care about the sampling distribution under the null hypothesis in Figure 5.3 but also about the sampling distribution under a competing alternative hypothesis. For simplicity, we first assume that we have two simple hypotheses, one representing the null (H_o) and one representing the alternative (H_1). Two simple hypothesis means that each can be represented by a single sampling distribution. We again summarize the results of individual studies into a single per-study Z-score (z_1 for the first study, z_2 for the second, etc). Under the null hypothesis, these Z-scores are generated by a normal distribution ϕ_0 with unit variance and mean 0; under the alternative hypothesis, these Z-scores are generated by an alternative distribution ϕ_1 with unit variance and mean 1.

The code in Figure 5.4 follows the same steps as the code in Figure 5.2 but it repeats each step for both both H_0 and H_1 in the `lapply` statements. We observe in Figure 5.5 that the same bias appears for the alternative hypothesis that we observe for the null hypothesis sampling distribution if we condition on arriving at our meta-analysis under extreme *Gold Rush* accumulation bias ($A(3) = 1$).

As a Bayesian, we simply do not care that our estimates are biased, as long as our posteriors are calibrated. We will first explain what calibration means for a Bayesian before we show that calibration stays intact under accumulation bias.

```
Hs <- c("H0" = "H0", "H1" = "H1")
numSim.study <- c("H0" = numSim.study, "H1" = numSim.study/10)  # sample ten times more H0 than H1
mean <- c("H0" = 0, "H1" = 1)

Z1 <- lapply(Hs, function(H) rnorm(numSim.study[H], mean = mean[H]))
Z2 <- lapply(Hs, function(H) rnorm(numSim.study[H], mean = mean[H]))
Z3 <- lapply(Hs, function(H) rnorm(numSim.study[H], mean = mean[H]))

# selection based on Gold Rush accumulation bias A(3) = 1
A3 <- lapply(Hs, function(H) which((Z1[[H]] > 1.96) & (Z2[[H]] > 1.96)))

# meta LRscores for a random sample of 3-study series (you don't need all for a smooth plot)
LRmeta3 <- lapply(Hs, function(H)
                  sapply(sample(1:numSim.study[H], size = numSim.study[H]/1000),
                         function(i) calcLRmeta(c(Z1[[H]][i], Z2[[H]][i], Z3[[H]][i]))))

# meta LRscores for a biased sample of 3-study series, biased by GoldRush A(3) = 1
LRmeta3.A3 <- lapply(Hs, function(H)
                     sapply(A3[[H]], function(i) calcLRmeta(c(Z1[[H]][i], Z2[[H]][i], Z3[[H]][i]))))

ggplot(rbind(data.frame(LR = LRmeta3[["H0"]],    H0H1 = "H0", GoldRush = ""),
             data.frame(LR = LRmeta3[["H1"]],    H0H1 = "H1", GoldRush = ""),
             data.frame(LR = LRmeta3.A3[["H0"]], H0H1 = "H0", GoldRush = "A3"),
             data.frame(LR = LRmeta3.A3[["H1"]], H0H1 = "H1", GoldRush = "A3"))) +
  geom_histogram(aes(x = LR,
                     y = ..density.., # ..density.. normalizes by group H0 or H1 with/without A(3) GoldRush
                     fill = interaction(H0H1, GoldRush)), # each with their own fill
                 bins = 120, position = "identity") +
  scale_x_continuous(trans = 'log10') +
  scale_fill_manual(values = hcl(c(15, 195, 15, 195), 100, 65, alpha = c(0.8, 0.8, 0.3, 0.3)),
                    name = bquote(H[0] ~ "or" ~ H[1] ~ "\n Conditioned on \n Gold Rush A(t)"),
                    labels = c(bquote(H[0] ~ ": " ~ LR^(3)),
                               bquote(H[1] ~ ": " ~ LR^(3)),
                               bquote(H[0] ~ ": " ~ LR^(3) ~ " |  A(3) = 1"),
                               bquote(H[1] ~ ": " ~ LR^(3) ~ " |  A(3) = 1")))
```

Figure 5.4. *Code to create Figure 5.5*

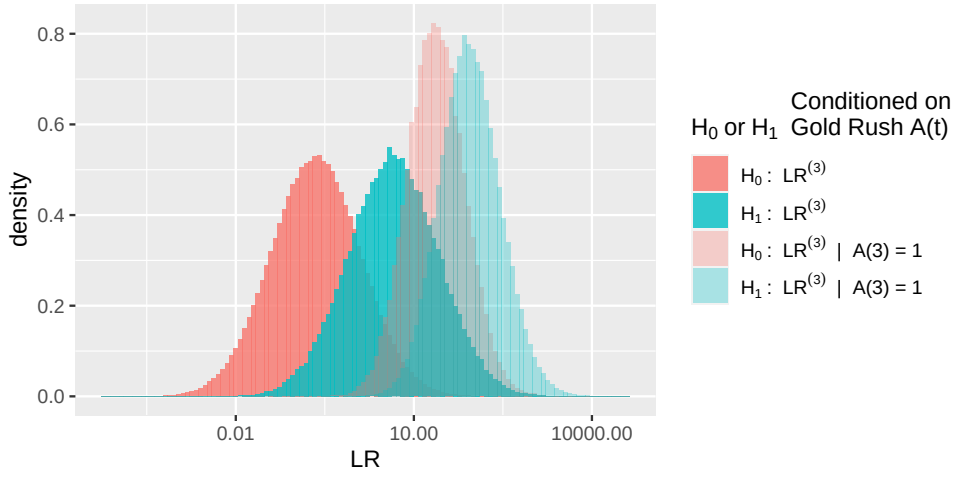

Figure 5.5. *Sampling distributions under the null (H_0) and alternative (H_1) hypothesis of likelihood ratios* $\mathbf{LR}^{(3)} = \prod_{i=1}^{3} \phi_0(Z_i)/\phi_0(Z_i)$ *of three studies with and without extreme* Gold Rush *accumulation bias* $A(t)$. *Note that the x-axis is on a log scale.*

Bayesian calibration of posterior odds and Bayesian error control

No accumulation bias To introduce the notion of Bayesian calibration of the posterior odds and Bayesian error control, we first turn to a situation without accumulation bias. Here we consider the posterior odds, but our discussion is closely related to the literature on Bayes factor calibration (De Heide and Grünwald, 2021; Hendriksen et al., 2020). We obtain the posterior odds by multiplying the likelihood ratio $\mathbf{LR}^{(3)}$ (`LRmeta3`) with a prior odds $\pi(H_1)/\pi(H_0)$ as follows:

$$\frac{\pi(H_1 \mid z_1, z_2, z_3)}{\pi(H_0 \mid z_1, z_2, z_3)} = \frac{\mathbf{P}(z_1, z_2, z_3 \mid H_1) \cdot \pi(H_1)}{\mathbf{P}(z_1, z_2, z_3 \mid H_0) \cdot \pi(H_0)} = \frac{\phi_1(z_1, z_2, z_3) \cdot \pi(H_1)}{\phi_0(z_1, z_2, z_3) \cdot \pi(H_0)} = \mathbf{LR}^{(3)} \cdot \frac{\pi(H_1)}{\pi(H_0)}.$$

```
gg.density <- ggplot(rbind(data.frame(postOdds = LRmeta3[["H0"]]*(1/10), H0H1 = "H0"),
                           data.frame(postOdds = LRmeta3[["H1"]]*(1/10), H0H1 = "H1"))) +
  geom_histogram(aes(x = postOdds,
                     y = ..density.., # ..density.. normalizes by group H0 or H1
                     fill = H0H1),
                 alpha = 0.8, bins = 120, position = "identity", show.legend = c(fill = FALSE)) +
  geom_vline(aes(xintercept = 16), linetype = "dashed") +
  scale_x_continuous(trans = 'log10',
                     breaks = c(0.01, 0.1, 1, 10, 100),
                     name = bquote(frac(pi(H[1]~"|"~z[1], z[2], z[3]),
                                        pi(H[0]~"|"~z[1], z[2], z[3])))))
gg.count <- ggplot(rbind(data.frame(postOdds = LRmeta3[["H0"]]*(1/10), H0H1 = "H0"),
                         data.frame(postOdds = LRmeta3[["H1"]]*(1/10), H0H1 = "H1"))) +
  geom_histogram(aes(x = postOdds,
                     y = ..count.., # ..count.. does not normalize by group H0 or H1
                     fill = H0H1), alpha = 0.8, bins = 120, position = "identity") +
  geom_vline(aes(xintercept = 16), linetype = "dashed") +
  scale_x_continuous(trans = 'log10',
                     breaks = c(0.01, 0.1, 1, 10, 100),
                     name = bquote(frac(pi(H[1]~"|"~z[1], z[2], z[3]),
                                        pi(H[0]~"|"~z[1], z[2], z[3])))) +
  scale_fill_discrete(name = "",
                      labels = c(bquote(H[0]), bquote(H[1])))
wrap_plots(list(gg.density, gg.count), ncol = 2, nrow = 1)
```

Figure 5.6. Code to create Figure 5.7

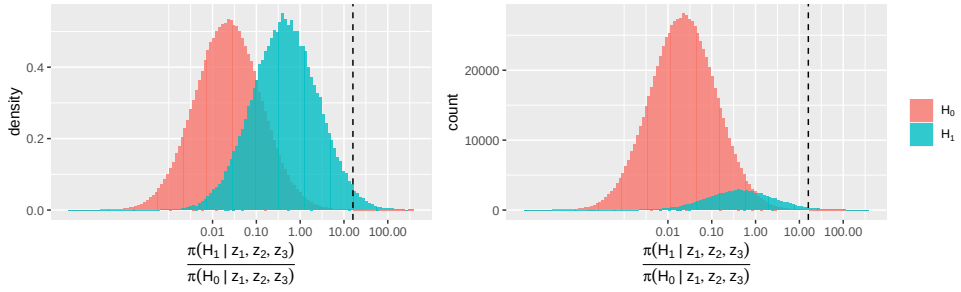

Figure 5.7. Sampling distributions under the null (H_0) and alternative (H_1) hypothesis of $\frac{\pi(H_1 \mid z_1, z_2, z_3)}{\pi(H_0 \mid z_1, z_2, z_3)} = \mathbf{LR}^{(3)} \cdot (1/10)$ with accumulation bias. The vertical line indicates the threshold for the posterior odds at $r = 16$. Note that the x-axis is on a log scale.

We take the unbiased sample of likelihood ratios in `LRmeta3` from the code Figure 5.4 and obtain the posterior odds (`postOdds`) in the code in Figure 5.6 for each likelihood ratio by multiplication with a prior odds odds of (1/10):

$$\frac{\pi(H_1 \,|\, z_1, z_2, z_3)}{\pi(H_0 \,|\, z_1, z_2, z_3)} = \mathbf{LR}^{(3)} \cdot \frac{\pi(H_1)}{\pi(H_0)} = \mathbf{LR}^{(3)} \cdot (1/10).$$

The result of this code is given by Figure 5.7 in two histograms for our sampled posterior odds. One using the statement `..density..` and one using `..count...` The first normalizes the histogram bars such that they add up to one. This is the same plot as Figure 5.5, just with the x-axis scaled by (1/10) because we show posterior odds instead of the likelihood ratio. The second histogram does something different: it just counts the samples and so the histogram bars scale with the number of samples we take in the `sample` statement in calculating `LRmeta3` in Figure 5.4.

Bayesian calibration With calibration of the Bayes posterior odds we mean that if we sample from both H_1 (which is ϕ_1 in our example) and H_0 (which is ϕ_0 in our example) and look at a posterior odds with value o_{post}, observing this value for the posterior odds makes (H_1) our alternative hypothesis o_{post} times more probable than (H_0) our null hypothesis. In other words: the posterior odds of obtaining posterior odds of o_{post} are o_{post}.

$$\frac{\mathbf{P}\left(H_1 \,\middle|\, \frac{\pi(H_1 \,|\, Z_1, Z_2, Z_3)}{\pi(H_0 \,|\, Z_1, Z_2, Z_3)} = o_{\text{post}}\right)}{\mathbf{P}\left(H_0 \,\middle|\, \frac{\pi(H_1 \,|\, Z_1, Z_2, Z_3)}{\pi(H_0 \,|\, Z_1, Z_2, Z_3)} = o_{\text{post}}\right)} = o_{\text{post}}$$

$$\text{because} \quad \frac{\mathbf{P}\left(\frac{\pi(H_1 \,|\, Z_1, Z_2, Z_3)}{\pi(H_0 \,|\, Z_1, Z_2, Z_3)} = o_{\text{post}} \,\middle|\, H_1\right)}{\mathbf{P}\left(\frac{\pi(H_1 \,|\, Z_1, Z_2, Z_3)}{\pi(H_0 \,|\, Z_1, Z_2, Z_3)} = o_{\text{post}} \,\middle|\, H_0\right)} \cdot \frac{\pi(H_1)}{\pi(H_0)} = o_{\text{post}}.$$

We can observe Bayesian calibration in the count plot in Figure 5.7, for example by looking at a posterior odds of 1.00 that has exactly the same count in both the histogram generated by H_1 and the one by H_0, which means that the ratio of counts is 1.00. This ratio of counts is calibrated, while the ratio of densities is not. The reason is that the ratio of densities does not take into account the prior odds: we take ten times as many samples from H_0 as from H_1 in the code in Figure 5.4. This agrees with the prior odds of (1/10) that we assume in calculating the posterior odds.

The ratio of densities gives:

$$\frac{\mathbf{P}\left(\frac{\pi(H_1 \,|\, Z_1, Z_2, Z_3)}{\pi(H_0 \,|\, Z_1, Z_2, Z_3)} = o_{\text{post}} \,\middle|\, H_1\right)}{\mathbf{P}\left(\frac{\pi(H_1 \,|\, Z_1, Z_2, Z_3)}{\pi(H_0 \,|\, Z_1, Z_2, Z_3)} = o_{\text{post}} \,\middle|\, H_0\right)},$$

while the ratio of counts gives:

$$\frac{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3)}{\pi(H_0 \mid Z_1,Z_2,Z_3)} \quad = \quad o_{\text{post}} \,\middle|\, H_1\right)}{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3)}{\pi(H_0 \mid Z_1,Z_2,Z_3)} \quad = \quad o_{\text{post}} \,\middle|\, H_0\right)} \cdot \frac{\pi(H_1)}{\pi(H_0)}.$$

Because we look at ratios, as we do if we look at odds, the scale of the counts in the figure does not matter. For simplicity, we are abusing notation a little bit and referring with probabilities **P** to densities, because our histograms of sampling distributions discretize our statistics in small intervals to give a probability instead of a density.

Bayesian error control From the calibration of the Bayes posterior odds we can obtain a notion of Bayesian error control as follows:

$$\frac{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3)}{\pi(H_0 \mid Z_1,Z_2,Z_3)} \quad \geq \quad r \,\middle|\, H_1\right)}{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3)}{\pi(H_0 \mid Z_1,Z_2,Z_3)} \quad \geq \quad r \,\middle|\, H_0\right)} \cdot \frac{\pi(H_1)}{\pi(H_0)} \quad \geq \quad r.$$

This Bayesian calibration is indeed the case for counts of $\mathbf{LR}^{(3)}$ (`LRmeta3`) in Figure 5.7 with the vertical dashed line $r = 16$. Figure 5.8 gives the calculation.

```
> sum(LRmeta3[["H1"]]*(1/10) > 16)/sum(LRmeta3[["H0"]]*(1/10) > 16)
[1] 31.59184
```

Figure 5.8. Code to show Bayesian error control for a threshold of $r = 16$ for the Bayes posterior odds.

$$\frac{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3)}{\pi(H_0 \mid Z_1,Z_2,Z_3)} \geq r \,\middle|\, H_1\right)}{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3)}{\pi(H_0 \mid Z_1,Z_2,Z_3)} \geq r \,\middle|\, H_0\right)} \cdot \frac{\pi(H_1)}{\pi(H_0)} = 31.59 \geq r = 16.$$

If we use r as a threshold to decide that we believe H_1 is true and H_0 is false, the probability that we make an error is r smaller than the probability that we are right. In other words: if we use r as a threshold for the posterior odds, the odds for a correct decision are at least r.

5.4 Bayesian error control under extreme *Gold Rush* accumulation bias

We can make the same plots under our scenario of extreme *Gold Rush* accumulation bias and observe calibration. In the count plot in Figure 5.7, the posterior odds of 1.0, for example, has exactly the same count in the histogram for H_0 as it has for H_1, which means that the ratio of counts is 1.0.

```
gg.density <- ggplot(rbind(data.frame(postOdds = LRmeta3[["H0"]]*(1/10), H0H1 = "H0"),
                           data.frame(postOdds = LRmeta3[["H1"]]*(1/10), H0H1 = "H1"))) +
  geom_histogram(aes(x = postOdds,
                     y = ..density.., # ..density.. normalizes by group H0 or H1
                     fill = H0H1),
                 alpha = 0.8, bins = 120, position = "identity", show.legend = c(fill = FALSE)) +
  geom_vline(aes(xintercept = 16), linetype = "dashed") +
  scale_x_continuous(trans = 'log10',
                     breaks = c(0.01, 0.1, 1, 10, 100),
                     name = bquote(frac(pi(H[1]~"|"~z[1], z[2], z[3]),
                                        pi(H[0]~"|"~z[1], z[2], z[3])))))
gg.count <- ggplot(rbind(data.frame(postOdds = LRmeta3[["H0"]]*(1/10), H0H1 = "H0"),
                         data.frame(postOdds = LRmeta3[["H1"]]*(1/10), H0H1 = "H1"))) +
  geom_histogram(aes(x = postOdds,
                     y = ..count.., # ..count.. does not normalize by group H0 or H1
                     fill = H0H1), alpha = 0.8, bins = 120, position = "identity") +
  geom_vline(aes(xintercept = 16), linetype = "dashed") +
  scale_x_continuous(trans = 'log10',
                     breaks = c(0.01, 0.1, 1, 10, 100),
                     name = bquote(frac(pi(H[1]~"|"~z[1], z[2], z[3]),
                                        pi(H[0]~"|"~z[1], z[2], z[3])))) +
  scale_fill_discrete(name = "",
                      labels = c(bquote(H[0]), bquote(H[1])))
wrap_plots(list(gg.density, gg.count), ncol = 2, nrow = 1)
```

Figure 5.9. *Code to create Figure 5.10*

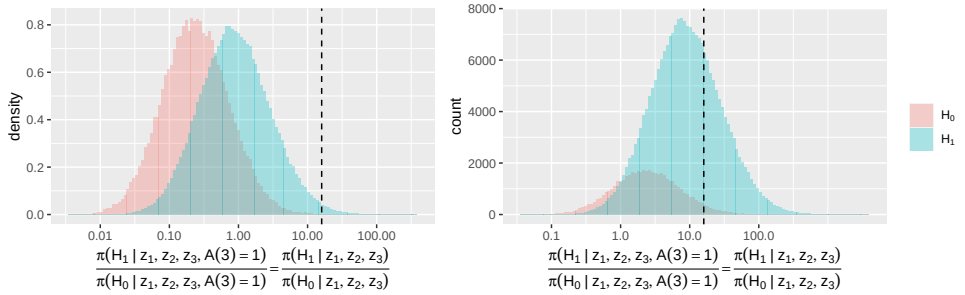

Figure 5.10. *Sampling distributions under the null (H_0) and alternative (H_1) hypothesis of* $\frac{\pi(H_1 \mid z_1,z_2,z_3)}{\pi(H_0 \mid z_1,z_2,z_3)} = \mathbf{LR}^{(3)} \cdot (1/10)$ *with accumulation bias. The vertical line indicates the threshold for the posterior odds at* $r = 16$*. Note that the x-axis is on a log scale.*

Now the ratio of densities gives:

$$\frac{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3,A(3)=1)}{\pi(H_0 \mid Z_1,Z_2,Z_3,A(3)=1)} \quad = \quad o_{\text{post}} \,\middle|\, A(3)=1,H_1\right)}{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3,A(3)=1)}{\pi(H_0 \mid Z_1,Z_2,Z_3,A(3)=1)} \quad = \quad o_{\text{post}} \,\middle|\, A(3)=1,H_0\right)},$$

while the ratio of counts gives:

$$\frac{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3,A(3)=1)}{\pi(H_0 \mid Z_1,Z_2,Z_3,A(3)=1)} \quad = \quad o_{\text{post}} \,\middle|\, A(3)=1,H_1\right)}{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3,A(3)=1)}{\pi(H_0 \mid Z_1,Z_2,Z_3,A(3)=1)} \quad = \quad o_{\text{post}} \,\middle|\, A(3)=1,H_0\right)} \cdot \frac{\mathbf{P}(A(3)=1 \mid H_1)}{\mathbf{P}(A(3)=1 \mid H_0)} \cdot \frac{\pi(H_1)}{\pi(H_0)}.$$

What counteracts the accumulation bias is illustrated in Figure 5.12. The posterior odds conditions on accumulating and analyzing three studies, $A(3) = 1$, which means that we are in a very biased sample. But because this situation occurs much more often under H_1 than under H_0

$$\mathbf{P}(A(3)=1 \mid H_1) >> \mathbf{P}(A(3)=1 \mid H_0),$$

our biased sample statistic $\mathbf{LR}^{(3)} \mid A(3) = 3$ can still achieve calibration if we take into account our prior odds.

In the ratio of counts, we also still have Bayesian error control under extreme *Gold Rush* accumulation bias:

```
> sum(LRmeta3.A3[["H1"]]*(1/10) > 16)/sum(LRmeta3.A3[["H0"]]*(1/10) > 16)
[1] 29.88396
```

Figure 5.11. *Code to show Bayesian error control for a threshold for the Bayes posterior odds of* $r = 16$.

$$\frac{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3,A(3)=1)}{\pi(H_0 \mid Z_1,Z_2,Z_3,A(3)=1)} \;\geq\; r \,\middle|\, A(3)=1,H_1\right)}{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3,A(3)=1)}{\pi(H_0 \mid Z_1,Z_2,Z_3,A(3)=1)} \;\geq\; r \,\middle|\, A(3)=1,H_0\right)} \cdot \frac{\mathbf{P}(A(3)=1 \mid H_1)}{\mathbf{P}(A(3)=1 \mid H_0)} \cdot \frac{\pi(H_1)}{\pi(H_0)} = 29.88 \geq r = 16.$$

We don't have to know the accumulation bias

How can it be that we never have to include anything about $A(3)$ in our calculations? The x-axis label of Figure 5.10 states that

$$\frac{\pi(H_1 \mid z_1,z_2,z_3,A(3)=1)}{\pi(H_0 \mid z_1,z_2,z_3,A(3)=1)} \quad = \quad \frac{\pi(H_1 \mid z_1,z_2,z_3)}{\pi(H_0 \mid z_1,z_2,z_3)},$$

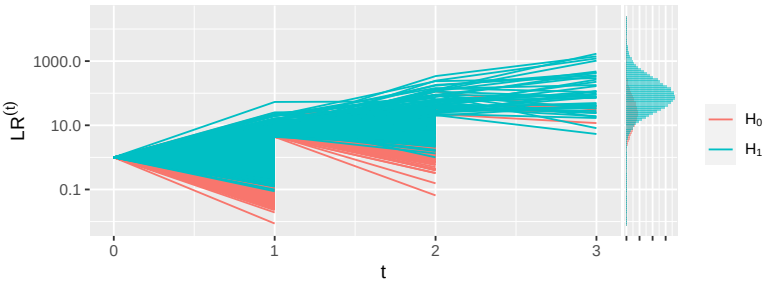

Figure 5.12. *Likelihood ratios* $\mathbf{LR}^{(1)}, \mathbf{LR}^{(2)}, \mathbf{LR}^{(3)}$ *when studies accumulate from 1 to 3 under the extreme* Gold Rush *accumulation bias process. Data simulated under prior odds* $H_1 : H_0 = 1 : 10$*. Note that the y-axis is logarithmic.*

```
numSim.study <- 5000  # these are not for the histogram, so a smaller simulation will do

Hs <- c("H0" = "H0", "H1" = "H1")
numSim.study <- c("H0" = numSim.study, "H1" = numSim.study/1)   # sample ten times more H0 than H1
mean <- c("H0" = 0, "H1" = 1)

Z1 <- lapply(Hs, function(H) rnorm(numSim.study[H], mean = mean[H]))
Z2 <- lapply(Hs, function(H) rnorm(numSim.study[H], mean = mean[H]))
Z3 <- lapply(Hs, function(H) rnorm(numSim.study[H], mean = mean[H]))

A1 <- lapply(Hs, function(H) 1:numSim.study[[H]])
A2 <- lapply(Hs, function(H) which(Z1[[H]] > 1.96))
A3 <- lapply(Hs, function(H) c("A3" = which((Z1[[H]] > 1.96) & (Z2[[H]] > 1.96))))

LR1 <- lapply(Hs, function(H) sapply(A1[[H]], function(i) calcLRmeta(Z1[[H]][i])))
LRmeta2.A2 <- lapply(Hs, function(H) sapply(A2[[H]], function(i) calcLRmeta(c(Z1[[H]][i], Z2[[H]][i]))))

# this is a smaller sample than LRmeta3.A3, which we use below for the histogram
LRmeta3.A3s <- lapply(Hs, function(H) sapply(A3[[H]], function(i) calcLRmeta(c(Z1[[H]][i], Z2[[H]][i], Z3[[H]][i]))))

ggplot(rbind(data.frame(t = 0, LR = 1,                    H0H1 = "H0", series = paste("H0", A1[["H0"]])),
             data.frame(t = 0, LR = 1,                    H0H1 = "H1", series = paste("H1", A1[["H1"]])),
             data.frame(t = 1, LR = LR1[["H0"]],          H0H1 = "H0", series = paste("H0", A1[["H0"]])),
             data.frame(t = 1, LR = LR1[["H1"]],          H0H1 = "H1", series = paste("H1", A1[["H1"]])),
             data.frame(t = 2, LR = LRmeta2.A2[["H0"]],   H0H1 = "H0", series = paste("H0", A2[["H0"]])),
             data.frame(t = 2, LR = LRmeta2.A2[["H1"]],   H0H1 = "H1", series = paste("H1", A2[["H1"]])),
             data.frame(t = 3, LR = LRmeta3.A3s[["H0"]],  H0H1 = "H0", series = paste("H0", A3[["H0"]])),
             data.frame(t = 3, LR = LRmeta3.A3s[["H1"]],  H0H1 = "H1", series = paste("H1", A3[["H1"]]))))) +
  geom_line(aes(x = t, y = LR, colour = H0H1, group = series)) +
  geom_ysidehistogram(aes(y = LR, x = ..count.., # ..count.. does not normalize by group H0 or H1
                          fill = H0H1),
                      data = rbind(data.frame(LR = LRmeta3.A3[["H0"]], H0H1 = "H0"),
                                   data.frame(LR = LRmeta3.A3[["H1"]], H0H1 = "H1")),
                      alpha = 0.8, bins = 100, position = "identity") +
  scale_y_continuous(trans = 'log10', name = bquote(LR^(t))) +
  scale_ysidex_continuous(labels = NULL) +
  scale_color_discrete(name = "", labels = c(bquote(H[0]), bquote(H[1])))
```

Figure 5.13. *Code to create Figure 5.12*

which follows because

$$\begin{aligned}
\frac{\pi(H_1 \,|\, z_1, z_2, z_3, A(3) = 1)}{\pi(H_0 \,|\, z_1, z_2, z_3, A(3) = 1)} &= \frac{\mathbf{P}(z_1, z_2, z_3, A(3) = 1 \,|\, H_1) \cdot \pi(H_1)}{\mathbf{P}(z_1, z_2, z_3, A(3) = 1 \,|\, H_0) \cdot \pi(H_0)} \\
&= \frac{\phi_1(z_1, z_2, z_3) \cdot A(3 \,|\, z_1, z_2, z_3) \cdot \pi(H_1)}{\phi_0(z_1, z_2, z_3) \cdot A(3 \,|\, z_1, z_2, z_3) \cdot \pi(H_0)} \\
&= \frac{\phi_1(z_1, z_2, z_3) \cdot \pi(H_1)}{\phi_0(z_1, z_2, z_3) \cdot \pi(H_0)} \\
&= \frac{\pi(H_1 \,|\, z_1, z_2, z_3)}{\pi(H_0 \,|\, z_1, z_2, z_3)} = \mathbf{LR}^{(3)} \cdot \frac{\pi(H_1)}{\pi(H_0)}.
\end{aligned}$$

This is the reason that given the sample values z_1, z_2, z_3, the only calculations we performed were to get $\mathbf{LR}^{(3)}$ in the `LRmeta3.A3` statement in Figure 5.2. We obtained our posterior odds in Figure 5.11 by simply multiplying $\mathbf{LR}^{(3)}$ with the prior odds $(1/10)$.

Given that we know the data, the probability of accumulating our studies is the same under the null and the alternative hypothesis and drops out of the ratio. We do not need to know what these are to calculate our posterior odds. What matters is how often we are in the null and the alternative situation: our prior odds. This is known as *stopping rule independence* (Hendriksen et al., 2020; Berger and Berry, 1988). If we know our prior odds, we do not need to know the accumulation bias process under which our study results z_1, z_2, z_3 were obtained (the probability $A(3 \mid z_1, z_2, z_3)$, see Chapter 3). We can just calculate our posterior odds in the usual way and decide on a threshold r on that posterior odds.

Prior odds

How often we reach $A(3) = 1$ under H_0 in comparison to under H_1 needs a statement of the relative occurrences of H_0 and H_1: a prior odds $\pi(H_1)/\pi(H_0)$. In the code in Figure 5.4 and Figure 5.13 we sample ten times as many null effects as alternative effects, so we assume that for every clinical trial that studies an effective treatment $\pi(H_1)$, we have $\pi(H_0)/\pi(H_1) = 10$ clinical trials that study an ineffective treatment, so $\pi(H_1)/\pi(H_0) = 1/10$. In Figure 5.12 we show that even if ten times as many studies observe data from the null hypothesis, still a lot more from the alternative hypothesis make it to a three-study-series under extreme *Gold Rush* accumulation bias.

5.5 The prior odds are crucial

The Bayesian calibration is driven by the fact that accumulation bias processes like the extreme *Gold Rush* make it much more likely for study series generated by H_1 to reach the meta-analysis than for study series generated by H_0. How much more depends on how many times either of them can try. As a Bayesian meta-analyst, we can think of this as a property of the research field that we might know and include in the analysis. How many initial studies are measuring a true effect from H_1 for each one that measures a null effect from H_0?

Bayesian calibration does rely crucially on getting the prior odds right. If we set our prior odds to a default $1:1$ and there is extreme *Gold Rush* accumulation bias in our field, we are actually assuming that hardly any series of clinical trials studying a null effect will accumulate three studies. Figure 5.14 shows what we are assuming in this case. For these plots we have set the following in the code in Figure 5.4 and Figure 5.13

```
numSim.study <- c("H0" = numSim.study, "H1" = numSim.study/1)
```

and we calculate the posterior odds based on our assumed $1:1$ prior odds:

$$\frac{\pi(H_1 \mid z_1, z_2, z_3, A(3) = 1)}{\pi(H_0 \mid z_1, z_2, z_3, A(3) = 1)} = \mathbf{LR}^{(3)} \cdot \frac{\pi(H_1)}{\pi(H_0)} = \mathbf{LR}^{(3)} \cdot \frac{1}{1} = \mathbf{LR}^{(3)}. \tag{5.1}$$

Figure 5.14 shows that assuming 1:1 prior odds in the calculations based on our data,

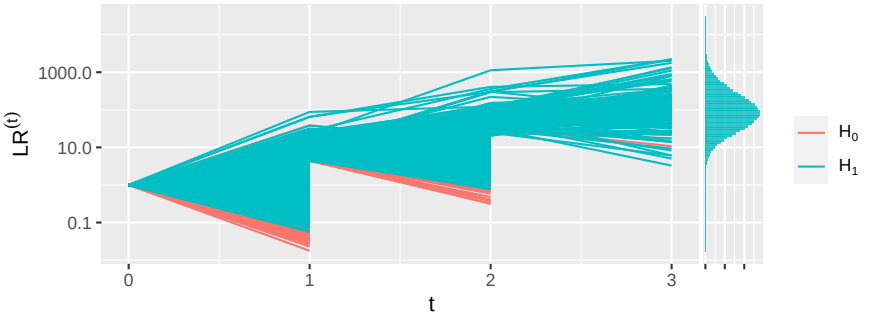

Figure 5.14. *Likelihood ratios* $\mathbf{LR}^{(1)}, \mathbf{LR}^{(2)}, \mathbf{LR}^{(3)}$ *when studies accumulate from 1 to 3 under the extreme* Gold Rush *accumulation bias process. Data simulated under prior odds* $H_1 : H_0 = 1 : 1$. *Note that the y-axis is logarithmic.*

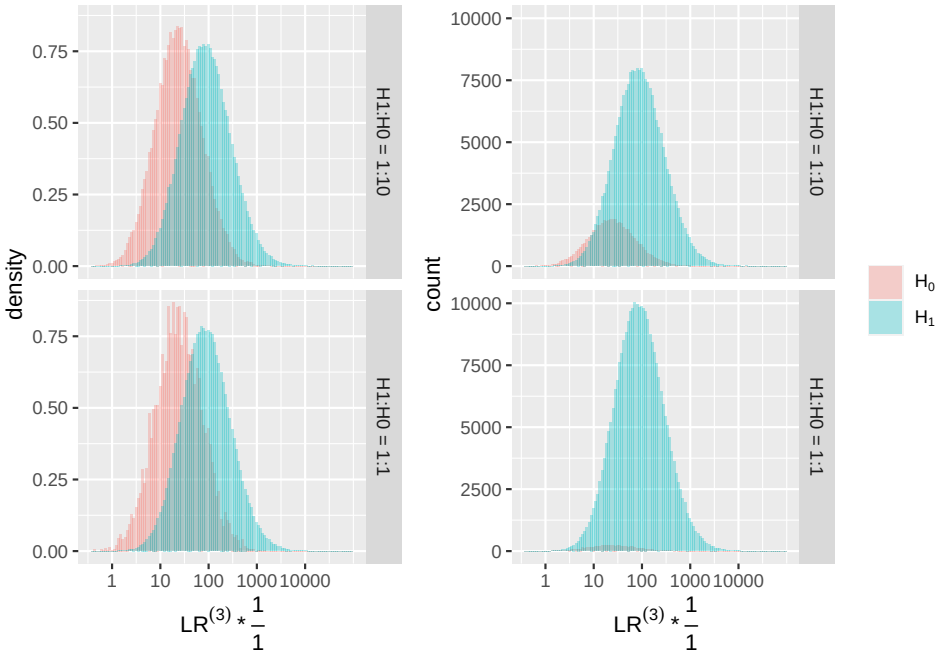

Figure 5.15. *Sampling distributions under the null* (H_0) *and alternative* (H_1) *hypothesis of* $\mathbf{LR}^{(3)} \cdot (1/1)$ *under extreme* Gold Rush *accumulation bias. The upper panels are sampled using* $H_1 : H_0 = 1 : 10$ *and the lower panels using* $H_1 : H_0 = 1 : 1$. *The upper right panel shows that mistakenly assuming* $1 : 1$ *in the posterior odds* $\mathbf{LR}^{(3)} \cdot (1/1)$ *does not give calibration under* $H_1 : H_0 = 1 : 10$, *e.g.* $\mathbf{LR}^{(3)} = 10$ *does not happen ten times as often under* H_0 *than under* H_1. *Note that the x-axis is on a log scale.*

when the true number of null hypotheses studied is ten times larger – the ratio in our research field is 1:10 – breaks the calibration of our posterior odds. For example, for an incorrectly calculated posterior odds $\mathbf{LR}^{(3)} \cdot (1/1) = 10$, we are in a situation that happens just as often under the null as under the alternative (the top-right panel of Figure 5.15) which should not give much evidence in favor of the alternative. As a result, also Bayesian error control breaks.

Setting prior odds is not that easy

We do want to stress that in the setting of retrospective meta-analysis, where the results of individual trials can be known to the meta-analyst before performing the analysis, it might be very difficult to establish prior odds that are not influenced by the data. In such scenarios, relying upon field-specific priors, e.g. established by prediction markets involving many peers (Potthoff, 2007; Dreber et al., 2015), might achieve more reliable prior odds.

What is more, these prior odds need to represent the ratio of alternative to null *initial studies* and not the ratio in meta-analyses. Reaching enough studies – e.g. $t = 3$ under extreme *Gold Rush* – and doing the meta-analysis is part of the data in the likelihood, not part of the prior. We encounter trouble with Bayesian calibration when we use different priors for individual studies than we use for a meta-analysis.

Doing so is appealing, though, since meta-analyses seem to be wrong less often than individual studies. The famous paper "Why Most Published Research Findings Are False" (Ioannidis, 2005b), for example, specifies different prior odds for a clinical trial analysis in comparison to a meta-analysis of clinical trials. This was the paper that introduced the concept of field-specific prior odds to a large audience as "Ratio of True to Not-True Relationships (R)". The different types of prior odds include one for "Adequately powered RCT with little bias" and one for "Confirmatory meta-analysis of good-quality RCTs". The first is set to an R of 1:1 and the second R to 2:1. This means that information seeped into the prior odds about what type of RCTs end up in meta-analyses; getting to the meta-analysis stage is assumed to be more likely under the alternative than the null otherwise the two prior odds would be the same. The meta-analysis prior odds that Ioannidis (2005b) specifies are essentially $\frac{\mathbf{P}(A(t)=1\,|\,H_1)}{\mathbf{P}(A(t)=1\,|\,H_0)} \cdot \frac{\pi(H_1)}{\pi(H_0)}$ with $\mathbf{P}(A(1) = 1\,|\,H_1) = 1$ and $\mathbf{P}(A(1) = 1\,|\,H_0) = 1$ for primary studies. This invalidates the stopping rule principle. Including information about the accumulation process into the prior biases the Bayes posterior and requires that the same information is included in the likelihood as well for Bayesian calibration. We need to know the accumulation bias process $A(t)$ in that case, which is usually impossible.

5.6 Beyond simple hypotheses

The situation with two simple hypotheses that we discussed so far, where each trial is either collecting data from ϕ_0 or ϕ_1, is not very realistic. More generally, we like to vary the parameter μ of our normal distribution ϕ_μ and allow for all possible normal distributions.

We assume we are given a *minimum relevant effect size* μ_{min} as well as a $\mu_0 < \mu_{\text{min}}$, which

respectively define the alternative hypothesis H_1 and the null hypothesis H_0:

$$H_0 = \{\phi_\mu : \mu \leq \mu_0\}, \quad H_1 = \{\phi_\mu : \mu \geq \mu_{\min}\}.$$

We can distinguish two types of prior probabilities: $\pi(H_1)$ and $\pi(H_0)$ for the hypotheses H_1, H_0, and $\pi_1(\mu)$ and $\pi_0(\mu)$ for $\{\mu : \mu \geq \mu_{\min}\}$ and $\{\mu : \mu \leq \mu_0\}$ respectively. Instead of a likelihood ratio of two simple hypotheses, we specify a Bayes Factor of two Bayes marginal distributions, using the priors on μ:

$$\mathbf{BF}(z_1, \ldots, z_t) = \frac{\bar{\phi}_1(z_1, \ldots, z_t)}{\bar{\phi}_0(z_1, \ldots, z_t)},$$

$$\text{with} \qquad \bar{\phi}_1(z) = \int \phi_\mu(z)\pi_1(\mu)dz \quad \text{and} \quad \bar{\phi}_0(z) = \int \phi_\mu(z)\pi_0(\mu)dz;$$

$$\bar{\phi}_1(z_1, \ldots, z_t) = \prod_{i=1}^{t} \bar{\phi}_1(z) \quad \text{and} \quad \bar{\phi}_0(z_1, \ldots, z_t) = \prod_{i=1}^{t} \bar{\phi}_0(z).$$

If π_j puts all its mass on a particular element μ^*, then $\bar{\phi}_j(z) = \phi_{\mu^*}(z)$.

Combining the Bayes Factor with the prior odds gives us the posterior odds, that just like the earlier posterior odds for two simple hypotheses, does not depend on the accumulation bias process for reaching e.g. $A(3) = 3$.

$$\frac{\pi(H_1 \mid z_1, z_2, z_3, A(3) = 1)}{\pi(H_0 \mid z_1, z_2, z_3, A(3) = 1)} = \frac{\bar{\phi}_1(z_1, z_2, z_3) \cdot A(3 \mid z_1, z_2, z_3)}{\bar{\phi}_0(z_1, z_2, z_3) \cdot A(3 \mid z_1, z_2, z_3)} \cdot \frac{\pi(H_1)}{\pi(H_0)} \tag{5.2}$$

$$= \frac{\bar{\phi}_1(z_1, z_2, z_3)}{\bar{\phi}_0(z_1, z_2, z_3)} \cdot \frac{\pi(H_1)}{\pi(H_0)} \tag{5.3}$$

$$= \mathbf{BF}(z_1, z_2, z_3) \cdot \frac{\pi(H_1)}{\pi(H_0)}. \tag{5.4}$$

The pseudo-Bayes posterior odds

What if we cannot come up with a good prior on the μs? In that case we may want to 'represent' the set of distributions H_0 and H_1 by their 'least extreme elements' respectively, i.e. ϕ_{μ_0} and $\phi_{\mu_{\min}}$. This gives the *pseudo-Bayes posterior odds*,

$$\frac{\pi^{\text{ps}}(H_1 \mid z_1, z_2, z_3)}{\pi^{\text{ps}}(H_0 \mid z_1, z_2, z_3)} = \frac{\phi_{\mu_{\min}}(z_1, z_2, z_3)}{\phi_{\mu_0}(z_1, z_2, z_3)} \cdot \frac{\pi(H_1)}{\pi(H_0)}$$

$$= \text{BF}^{\text{ps}}(z_1, z_2, z_3) \cdot \frac{\pi(H_1)}{\pi(H_0)}.$$

which is just the 'real' posterior odds (5.2) that we would get if we had put all our prior mass on $\mu_{\min}$ and μ_0 respectively.

We can use the 'pseudo-Bayes posterior odds' when (with a Bayesian mindset) we have no good idea about 'good' priors on the μs or (with a frequentist mindset) about what

value of μ may be true if H_1 is true, or what value of μ may be true if H_0 is true. Note in particular that in the pseudo-Bayes posterior odds, we use the *same* priors on H_0 and H_1 as in the 'real' posterior, but different, degenerate priors on the μs.

The GROW e-values that we calculate in ALL-IN meta-analysis (Chapter 1) are pseudo-Bayes factors BF^{ps}. So in ALL-IN meta-analysis, we can very simply extend our conclusions with Bayesian statements by combining our e-values with prior odds to obtain pseudo-Bayes posterior odds. Moreover, with these pseudo-Bayes posterior odds, we can also obtain Bayesian error control.

5.7 Pseudo-Bayesian error control

Throughout this blog post we have shown that if we get the prior odds right, the posterior odds is calibrated under accumulation bias, i.e.:

$$\frac{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3,A(3)=1)}{\pi(H_0 \mid Z_1,Z_2,Z_3,A(3)=1)} \;=\; o_{\text{post}} \,\middle|\, A(3)=1, H_1\right)}{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3,A(3)=1)}{\pi(H_0 \mid Z_1,Z_2,Z_3,A(3)=1)} \;=\; o_{\text{post}} \,\middle|\, A(3)=1, H_0\right)} \cdot \frac{\mathbf{P}(A(3)=1 \mid H_1)}{\mathbf{P}(A(3)=1 \mid H_0)} \cdot \frac{\pi(H_1)}{\pi(H_0)}$$

$$= \frac{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3)}{\pi(H_0 \mid Z_1,Z_2,Z_3)} \;=\; o_{\text{post}} \,\middle|\, A(3)=1, H_1\right)}{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3)}{\pi(H_0 \mid Z_1,Z_2,Z_3)} \;=\; o_{\text{post}} \,\middle|\, A(3)=1, H_0\right)} \cdot \frac{\mathbf{P}(A(3)=1 \mid H_1)}{\mathbf{P}(A(3)=1 \mid H_0)} \cdot \frac{\pi(H_1)}{\pi(H_0)} = o_{\text{post}}$$

$$\text{such that} \qquad \frac{\mathbf{P}\left(H_1 \,\middle|\, \frac{\pi(H_1 \mid Z_1,Z_2,Z_3)}{\pi(H_0 \mid Z_1,Z_2,Z_3)} \;=\; o_{\text{post}}, \quad A(3)=1\right)}{\mathbf{P}\left(H_0 \,\middle|\, \frac{\pi(H_1 \mid Z_1,Z_2,Z_3)}{\pi(H_0 \mid Z_1,Z_2,Z_3)} \;=\; o_{\text{post}}, \quad A(3)=1\right)} = o_{\text{post}}.$$

The outer probabilities combine a prior odds with a likelihood ratio of the observing the posterior odds of o_{post}, which is a statistic of our data, with the likelihood ratio of observing a three study series ($A(3)=1$), also part of the data. The likelihood ratio of the data combined with prior odds forms posterior odds for the hypotheses conditioned on the data.

We can use this fact of calibration to specify the Bayesian error control further for the pseudo-Bayes posterior odds. We define a threshold on the pseudo-Bayes posterior odds r and decide to reject the null hypothesis and believe the alternative if

$$\frac{\pi^{\text{ps}}(H_1 \mid z_1,\ldots,z_t)}{\pi^{\text{ps}}(H_0 \mid z_1,\ldots,z_t)} \geq r.$$

We can set a threshold such that if we cross it, we reject the null hypothesis and denote so by $\text{REJECT}[A(t)=1, r]$ based on crossing the threshold with our pseudo-Bayes posterior odds conditioned on accumulating t studies.

For a subset of all accumulation bias processes $A(t)$ which includes the extreme *Gold rush* and variations of it, we have the following: for all $t = 1,2,\ldots$ $r > 1$: the *true* Bayes

posterior odds – so not only the pseudo-Bayes posterior odds! – of H_0 satisfies:

$$\frac{\pi(H_1 \mid \text{REJECT}[A(t) = 1, r])}{\pi(H_0 \mid \text{REJECT}[A(t) = 1, r])} \geq r. \tag{5.5}$$

This expresses that, as long as the priors on H_0 and H_1 are chosen correctly, we have Bayesian error control for the pseudo-Bayes posterior odds: a Bayesian's real posterior odds of an incorrect decision can be no larger than the odds to make an error according to the pseudo-Bayes posterior odds on $\{H_0, H_1\}$, even though that the priors on the μs are, according to that same Bayesian, incorrect. Note that this is merely a 'one-sided' calibration, but the inequalities go the right (i.e. practically useful) way. The result holds not just for the normal location family but for a general class of models including all 1-dimensional exponential families and the 1-sample t-test setting. This general result is stated and proved in Appendix Section 5.A and Section 5.B.

5.8 Conclusion

In our imaginary world of extreme *Gold Rush* accumulation bias, the sampling distribution of the meta-analysis Z-score behaves very different from the sampling distribution assumed to calculate p-values and confidence intervals. A meta-analysis sampling distribution conditions on the available number of studies, which means that we are in a situation that is influenced by a selection effect: only some series get there, not others. Bayesian analysis also conditions on the number of studies and therefore also likelihood ratios and Bayes factors have biased sampling distributions. For Bayesian error control, however, we do not need unbiased sampling; we need *calibration*.

Some accumulation bias is at play in almost any retrospective meta-analysis. Bayesian calibration holds no matter the accumulation bias process, as long as the prior odds are trustworthy. The e-values from ALL-IN meta-analysis can also be used to combine with prior odds and obtain pseudo-Bayes posterior odds to obtain Bayesian error control through calibration, although now calibration may only hold for a subset of all accumulation bias processes – which however include the extreme *Gold Rush* scenario. This also allows for continuous monitoring; multiple testing is no problem, as long as the prior odds are correct. Setting trustworthy prior odds in meta-analysis is not easy, however. As long as the prior odds can be trusted, a Bayesian perspective on meta-analysis will reduce research waste by allowing efficient data-driven decisions – not letting them invalidate the inference – and still analyze the posterior odds for any particular meta-analysis, conditioned on arriving at the number of studies so far.

Code availability

This blogpost's R code is available on `https://osf.io/p2rtw/` (Ter Schure, 2021a).

Appendices

5.A Pseudo-Bayes posterior odds for exponential families and beyond

5.A.1 Exponential families

Let $\mathscr{M} = \{P_\delta : \delta \in \Delta\}$ with Δ a (possibly unbounded) open interval in $\mathbb{R}$ represent a 1-dimensional exponential family of probability distributions for some random variable Y, given in its mean-value parameterization. While this already includes important models such as the normal location family (z-test), the Bernoulli model, the Poisson model, and so on, we will later, in Theorem 5.B.1, extend our result to some multi-dimensional families that are not of exponential form. Each P_δ has a density (for continuous-valued Y) or mass function (for discrete-valued Y) p_δ. P_δ and p_δ are extended to i.i.d. sequences by taking product distributions.

We assume we are given two parameter values δ^- and δ^+ in Δ with $\delta^- < \delta^+$ which respectively define the *alternative* hypothesis H_1 and the null hypothesis H_0:

$$H_0 = \{P_\delta : \delta \in \Delta_0\},\ \Delta_0 = \{\delta \in \Delta : \delta \leq \delta^-\},\ H_1 = \{P_\delta : \delta \in \Delta_1\},\ \Delta_1 = \{\delta \in \Delta : \delta \geq \delta^+\}.$$

In the case treated in the main text, $\mathscr{M}$ denotes the normal location family, $\delta^+ = \mu_{\min}$ is the minimum clinically relevant effect size, and $\delta^- = \mu_0$. The fact that the treatment below is entirely symmetric in δ^+ and δ^- (if we swap δ^+ and δ^- and take the reciprocal of all Bayes factors and posterior odds, we get the same result) motivates the switch of notation. Still, in practice we will often have δ^+ interpretable as minimal effect size, $\Delta = \mathbb{R}_0^+$ and $\delta^- = 0$. We need the concept of a *stopping time*. We define this in a standard way in terms of a *randomized stopping rule*. This is any function f from outcome sequences of arbitrary length to $[0, 1]$. The interpretation is that for any actually generated initial sequence of data $y^n = y_1, y_2, \ldots, y_n$, we toss an independent coin with bias $f(y^n)$ after having observed y^n. If the coin lands tails, we stop. If not, we generate y_{n+1} and repeat the procedure for $n + 1$, etc. The stopping time τ is then the random variable set equal to the smallest n at which the coin has landed tails. *Gold Rush* accumulation bias (Chapter 3) defines such a stopping rule. The extreme version of *Gold Rush* accumulation bias presented in this blog post is a non-randomized version of this stopping rule.

5.A.2 The Bayes and the pseudo-Bayes posterior

We can distinguish two types of prior probabilities: $\pi(H_1)$ and $\pi(H_0)$ for the hypotheses H_0, H_1, and Π_1 and Π_0 for the parameter spaces $\Delta_1 = \{\delta : \delta \geq \delta^+\}$ and $\Delta_0 = \{\delta : \delta \leq \delta^-\}$. Π_j can be interpreted as the prior of δ conditioned on it lying in Δ_j; in general, we will allow priors that do not restrict δ to lie in Δ_j (i.e. we may have $\Pi(\delta \notin \Delta_0 \cup \Delta_1) = \pi' > 0$ and then $\pi(H_1) + \pi(H_0) + \pi' = 1$). We can calculate conditional and posterior probabilities and odds in the standard way using Bayes' theorem: illustrating a particular case we need later on, for general measurable events $\mathscr{E}_1$,

$$\frac{\pi(H_1 \mid \mathscr{E}_1)}{\pi(H_0 \mid \mathscr{E}_1)} = \frac{\bar{P}_1(\mathscr{E}_1)}{\bar{P}_0(\mathscr{E}_1)} \cdot \frac{\pi(H_1)}{\pi(H_0)}, \tag{5.A.1}$$

where $\bar{P}_j$ is the Bayes marginal distribution based on the prior Π_j. If Π_j has density π_j, then $\bar{P}_j(\mathscr{E}) = \int P_\delta(\mathscr{E})\pi_j(\delta)$. If Π_j puts all its mass on a particular element δ^*, then $\bar{P}_j(\mathscr{E}) = P_{\delta^*}(\mathscr{E})$.

In our setting, we observe a sequence $y_1, \ldots, y_n$, where n is itself the value that the stopping time τ (whose general underlying definition in terms of some stopping rule f may be unknown to us) takes; so we really observe $Y^n = y^n; \tau = n$. Because of the well-known fact that the Bayes posterior does not depend on the definition of the stopping time as long as it is defined in the (standard) way above (Hendriksen et al., 2020), we have for all n that $\pi(H_1 \mid Y^n = y^n, \tau = n) = \pi(H_1 \mid Y^n = y^n)$. i.e. if a variable stopping time is used and we happen to stop at $\tau = n$, the posterior is the same as if the sample size had been fixed in advance to n. This allows to express the *Bayes factor* $\mathbf{BF}(Y^\tau)$ and the posterior odds $\pi(H_1 \mid Y^\tau)/\pi(H_0 \mid Y^\tau)$ compactly as follows:

$$\bar{p}_j(Y^\tau) = \int_{\delta \in \Delta^+} p_\delta(Y^\tau) d\pi_j(\delta) \ , \text{ for all } n, \text{ for } j \in \{0, 1\} \ .$$
$$\mathbf{BF}(Y^\tau) = \frac{\bar{p}_1(Y^\tau)}{\bar{p}_0(Y^\tau)} \ , \ \frac{\pi(H_1 \mid Y^\tau)}{\pi(H_0 \mid Y^\tau)} = \mathbf{BF}(Y^\tau) \cdot \frac{\pi(H_1)}{\pi(H_0)}.$$

where we again assume that Π_j has density π_j; again, if Π_j puts all its mass on a particular element δ^*, then $\bar{p}_j(Y^\tau) = p_{\delta^*}(Y^\tau)$.

The pseudo-Bayes posterior odds What if we cannot come up with a good prior on Δ_0 and/or Δ_1? In that case we may want to 'represent' the set of distributions H_0 and H_1 by their 'least extreme elements' respectively, i.e. P_{δ^-} and P_{δ^+}. This gives the *pseudo-Bayes posterior odds* (given $\mathscr{E}_1, \mathscr{E}_2$ based on a prior that was already conditioned on $\mathscr{E}_1$)

$$\frac{\pi^{\text{ps}}(H_1 \mid \mathscr{E}_1, \mathscr{E}_2)}{\pi^{\text{ps}}(H_0 \mid \mathscr{E}_1, \mathscr{E}_2)} = \frac{P_{\delta^+}(\mathscr{E}_1 \mid \mathscr{E}_2)}{\bar{P}_{\delta^-}(\mathscr{E}_1 \mid \mathscr{E}_2)} \cdot \frac{\pi(H_1 \mid \mathscr{E}_2)}{\pi(H_0 \mid \mathscr{E}_2)}, \tag{5.A.2}$$

which is just the 'real' posterior odds that we would get if we had put all our prior mass on δ^+ and δ^- respectively. Similarly we get the *pseudo-Bayes factor*

$$\text{BF}^{\text{ps}}(Y^\tau) = \frac{p_{\delta^+}(Y^\tau)}{p_{\delta^-}(Y^\tau)}.$$

We can use the 'pseudo-Bayes factor' when (with a Bayesian mindset) we have no good idea about what might be a 'good' prior conditioned on $\delta \in \Delta^+$ and/or conditioned on $\delta \in \Delta^-$ or (with a frequentist mindset) about what value of δ in Δ^+ may be true if H_1 is true, or what value of δ in Δ^- may be 'true' if H_0 is true. Based on the pseudo-Bayes factor, we can also calculate the *pseudo-Bayes posterior odds* as if the Bayes factor were correct:

$$\frac{\pi^{\text{ps}}(H_1 \mid Y^\tau)}{\pi^{\text{ps}}(H_0 \mid Y^\tau)} = \text{BF}^{\text{ps}}(Y^\tau) \cdot \frac{\pi(H_1)}{\pi(H_0)}. \tag{5.A.3}$$

In case $\pi(H_0)+\pi(H_1)=1$, the pseudo-Bayes posterior probability of H_0 is given by:

$$\pi^{\text{ps}}(H_0 \mid Y^\tau) = \frac{p_{\delta^-}(Y^\tau)\cdot\pi(H_0)}{p_{\delta^+}(Y^\tau)\cdot\pi(H_1)+p_{\delta^-}(Y^\tau)\cdot\pi(H_0)} = \frac{1}{\text{BF}^{\text{ps}}(Y^\tau)\cdot(\pi(H_1)/\pi(H_0))+1}.$$

Note in particular that in the pseudo-Bayes posterior, we use the *same* priors on H_0 and H_1 as in the 'real' posterior, but different, degenerate priors on Δ_0 and Δ_1.

5.A.3 The Result

Fix a *significance threshold* $r>1$ and let τ be an arbitrary stopping time. We will *reject* H_0 if $\pi^{\text{ps}}(H_1 \mid Y^\tau)/\pi^{\text{ps}}(H_1 \mid Y^\tau) \geq r$, and *accept* H_0 otherwise. Let $\text{REJECT}_{\tau,r}$ be the event that we reject at level r when using stopping time τ (importantly, in practice it may be unknowable what stopping rule τ is actually being used; to calculate posterior probabilities we only need to know the observed data Y^τ and the sample size of the observed data, and not the general definition of τ, i.e. we do not need to know if we would have stopped at the same n if the data had been different).

Theorem 5.A.1. *Let $\{P_\delta : \delta\in\Delta\}$ represent a 1-dimensional exponential family as above. Fix some $\delta^-<\delta^+$ and define H_1, H_0 and BF^{ps} correspondingly; also fix some arbitrary priors $\pi(H_0), \Pi_0$ on Δ_0, Π_1 on Δ_1. We have the following: for each n and each $r>1$ and each stopping time τ such that*

$$\frac{\bar{P}_1(\tau=n)}{P_{\delta^+}(\tau=n)}\cdot\frac{P_{\delta^-}(\tau=n)}{\bar{P}_0(\tau=n)} \geq 1, \tag{5.A.4}$$

we have: the posterior odds given rejection at time n are well-defined and satisfy

$$\frac{\pi\left(H_1 \mid \text{REJECT}_{\tau,r}, \tau=n\right)}{\pi\left(H_0 \mid \text{REJECT}_{\tau,r}, \tau=n\right)} \geq r. \tag{5.A.5}$$

The theorem implies the statement (5.5) in the main text for the Gaussian location family. The t there corresponds to the n here, and the statement $A(t)=1$ to $\tau=n$; the process $A(t)$ defines the stopping rule τ. The required condition (5.A.4) is easily seen to hold for the *Gold rush* scenario in which we evaluate invariably at $t(=n)=3$: inspecting the definition of $A(t)$, we see that the probability of reaching time 3 (i.e. $A(3)=1$) under P_δ increases monotonically with δ if δ represents the mean of a normal distribution. $\bar{P}_1$ being a mixture of P_δ's with $\delta\geq\delta^+$ and $\bar{P}_0$ being a mixture of P_δ's with $\delta\leq\delta^0$, (5.A.4) then follows.

In case $\pi(H_0)+\pi(H_1)=1$ (we rule out that δ does not lie in $\Delta_0\cup\Delta_1$), we can alternatively work on the scale of probabilities rather than probability ratios and fix a significance level $0<\alpha<1/2$ and reject H_0 if $\pi^{\text{ps}}(H_0 \mid Y^\tau)\leq\alpha$. This is equivalent to the event $\text{reject}_{\tau,r_\alpha}$ with $r_\alpha=(1-\alpha)/\alpha$. (5.A.5) then expresses that for each $0<\alpha<1/2$

$$\pi\left(H_0 \mid \text{REJECT}_{\tau,r_\alpha}\right) \leq \alpha. \tag{5.A.6}$$

The theorem expresses that, as long as the priors on H_0 and H_1 are chosen correctly, the error probabilities of decisions on the pseudo-Bayes posterior are *calibrated*: a Bayesian's real posterior odds of the decision 'reject' being correct (given by conditioning the true prior on the observed stopping time and the fact that at that stopping time, we rejected) can be no smaller than the posterior odds that this decision is correct according to the pseudo-Bayes posterior distribution on $\{H_0, H_1\}$, even though that distribution is, according to that same Bayesian, incorrect. Note that this is merely a 'one-sided' calibration, but the inequalities go the right (i.e. practically useful) way.

In a more frequentist interpretation, we may think of $\pi(H_0)$ as the 'population frequency' that the null is true in the particular field of science that we are working in. Whenever in a study H_0 is true, a particular $\delta_0 \in \Delta_0$ will be 'true' and generate the data , and whenever in a study H_1 is true, a particular $\delta_1 \in \Delta_1$ will be 'true' and generate the data. Theorem 5.A.1 expresses that our *conditional* error probability for rejecting/accepting H_0 is smaller than α, even though we do not know the true δ_0 and δ_1's.

From both a Bayesian and a frequentist stance, the result says that as long as *our prior* $\pi(H_0)$ *on* H_0 *reflects what happens in the real world* and we use if for reject/accept decisions of the kind above, it is o.k. to use the pseudo-Bayes posterior, and we can get away with not having a 'correct' or 'better' prior on the parameters in Δ^+ and Δ^-.

5.B Extension and Proof of Theorem 5.A.1

Our result holds more generally than for i.i.d. exponential families. Namely, we can more generally let $\mathcal{M} = \{P_{\delta,\gamma} : \delta \in \Delta, \gamma \in \Gamma\}$ with Δ a (possibly unbounded) interval in $\mathbb{R}$ denote a family of distributions for some random process $U_1, U_2, \ldots$. Again, δ denotes the 1-dimensional parameter of interest (e.g. an effect size) and now γ denotes potential nuisance parameters. We assume again a δ^+ and a $\delta^- < \delta^+$ are given, defining the null and alternative hypotheses

$$H_0 = \{P_{\delta,\gamma} : \delta \leq \delta^-, \gamma \in \Gamma\} \quad H_1 = \{P_{\delta,\gamma} : \delta \geq \delta^+, \gamma \in \Gamma\}.$$

Our result is valid for general families of this form, if furthermore the following holds: there exists a sequence of random vectors $Y_1, Y_2, \ldots$ such that Y_n is determined (can be written as a function of) $U^n = (U_1, \ldots, U_n)$ and the following two properties hold:

Irrelevance of γ and Full Support The distribution $P_\delta^{(n)}$ of Y^n under process $P_{\delta,\gamma}$ is the same for all γ (hence we can omit it from the notation in $P_\delta^{(n)}$). It has a density $p_\delta^{(n)}$ relative to some fixed underlying measure, and this density has the same support for all $\delta \in \Delta$. That is, we require for all $y^n \in \mathbb{R}^n$ that if for some $\delta \in \Delta$, $p_\delta^{(n)}(y^n) > 0$, then for all $\delta \in \Delta$, $p_\delta^{(n)}(y^n) > 0$. As a consequence, for any stopping rule τ, for every n, if for some $\delta \in \Delta$ we have $P_\delta(\tau = n) > 0$ then for all $\delta \in \Delta$ we have $P_\delta(\tau = n) > 0$. We call the set of n with $P_\delta(\tau = n) > 0$ the *support* of τ.

Monotone likelihood ratio (MLR) Property There exists a function s_n on $\mathbb{R}^n$ such that for each $\delta_0 < \delta_1$ with $\delta_0, \delta_1 \in \Delta$, the likelihood ratio $\frac{p_{\delta_1}^{(n)}(Y^n)}{p_{\delta_0}^{(n)}(Y^n)}$ is an increasing func-

tion of random variable $S_n := s_n(Y^n)$.

Note that both properties automatically hold for 1-dimensional i.i.d. exponential families as above – then we can set Γ to be a singleton, then γ plays no role, we can take $Y_n = U_n$ and $S_n = s_n(Y^n)$ to be the sufficient statistic for n outcomes (if Y_1 is the sufficient statistic for one outcome, then $S_n = \sum_{i=1}^n Y_i$) and then both properties are easily verified (Lehmann, 1986). But they also hold in the t-test setting, where $P_{\delta,\gamma}$ states that the underlying data U_i are i.i.d. normally distributed with variance γ and effect size δ (i.e. mean $\mu = \delta\gamma$). We can then take $Y_i := U_i/|U_1|$ to be the so-called 'maximal invariant statistic' (Hendriksen et al., 2020) and S_n to be the t-statistic based on U^n, which can be written as a function of Y^n. It is a well-known fact that S_n has a non-central t-distribution and that this satisfies the MLR property (Lehmann, 1986). We note that the set of allowed stopping rules/times remains unchanged in this more general set-up. Thus, in the t-test setting, and more generally in settings with $U_i \neq Y_i$, the stopping rule $f(Y^n)$ at time n must be writeable as a function of the Y^n which is a coarsening of (contains less information than) the data U^n. Since the condition above implies that the likelihood ratio can be written as a function of Y^n, and we usually use stopping rules that depend on the likelihood ratio observed so far and possibly some additional data that is independent of the observed data, but nothing else, this poses no great restriction in practice.

We now formulate and prove the theorem for this more general setup. Generalizing (5.A.3), the pseudo-Bayes posterior odds are now defined as:

$$\frac{\pi^{\text{ps}}(H_1 \mid Y^\tau)}{\pi^{\text{ps}}(H_0 \mid Y^\tau)} = \text{BF}^{\text{ps}}(Y^\tau) \cdot \frac{\pi(H_1)}{\pi(H_0)} \text{ with } \text{BF}^{\text{ps}}(Y^\tau) = \frac{p^{(\tau)}_{\delta^+}(Y^\tau)}{p^{(\tau)}_{\delta^-}(Y^\tau)}. \tag{5.B.1}$$

Let again $\text{REJECT}_{\tau,r}$ be the event that $\pi^{\text{ps}}(H_1 \mid Y^\tau)/\pi^{\text{ps}}(H_1 \mid Y^\tau) \geq r$.

Theorem 5.B.1. *Let $\{P_{\delta,\gamma} : \delta \in \Delta, \gamma \in \Gamma\}$ represent a family that satisfies the two properties above for all n. Fix some $\delta^- < \delta^+$ and define H_1, H_0 and BF^{ps} correspondingly; also fix some arbitrary priors $\pi(H_0), \pi(H_1)$ and Π_0 on Δ_0, Π_1 on Δ_1. We have the following for each $r > 1$ and each stopping time τ and each n in the support of τ: the* true *posterior odds of H_1 vs. H_0 satisfy:*

$$\frac{\pi\left(H_1 \mid \text{REJECT}_{\tau,r}, \tau = n\right)}{\pi\left(H_0 \mid \text{REJECT}_{\tau,r}, \tau = n\right)} \geq r \cdot \frac{\bar{P}_1(\tau = n)}{\bar{P}_0(\tau = n)} \cdot \frac{P_{\delta^-}(\tau = n)}{P_{\delta^+}(\tau = n)}. \tag{5.B.2}$$

The earlier Theorem 5.A.1 is immediately seen to be a special case.

Remark The fact that we can go beyond exponential families raises the question of how general the result is. In this respect, we note that our conditions imply that the sequence of pseudo-Bayes factors $\text{BF}^{\text{ps}}(Y^1), \text{BF}^{\text{ps}}(Y^2), \ldots$ in (5.B.1) define a *test martingale* or equivalently, a product of conditional *E-values* under H_0 (Grünwald et al., 2019). Interestingly, *unconditional* frequentist error control under arbitrary stopping times can be given for arbitrary test martingales. All Bayes factors satisfying the conditions of the general version of the theorem below define *two-sided* test martingales: by this, we mean

that $1/\text{BF}^{\text{ps}}(Y^1), 1/\text{BF}^{\text{ps}}(Y^2), \ldots$ defines a test martingale under H_1. One might therefore suspect that our result continues to hold whenever we set our pseudo-Bayes factor equal to a two-sided test martingale, even if the MLR property does not hold. But it is not clear whether this really is the case. An example is the safe logrank test of Ter Schure et al. (2020b) (Chapter 2). The pseudo-Bayes factor we develop there is a ratio of partial likelihoods, and it defines a two-sided test martingale. Nevertheless, it is easily seen that due to the data not being i.i.d. the MLR property does *not* hold, and this property seems crucial for the argument used in the proof. Whether or not a (perhaps slightly weakened, i.e. $\geq r$ in (5.A.6) replaced by $\geq cr$ for some $c < 1$) version of the theorem holds for general pseudo-Bayes factors given by general two-sided test martingales is an interesting topic for future research.

Proof of Theorem 5.B.1

Fix $n \in \mathbb{N}$ in the support of τ. The proof makes crucial use of Lemma 5.B.2, which we state and prove first. We prove the theorem and the lemma only for the discrete case (with each Y_i taking values in a countable set $\mathcal{Y}_i \subset \mathbb{R}$), for which all densities become probability mass functions. It is straightforward to extend the results to the general case by replacing all probability mass functions with appropriate densities and sums by integrals.

Lemma 5.B.2. *Suppose that the MLR Property holds for some given n in the support of τ relative to some S_n as above. Then*

1. *The MLR Property holds for the set of distributions $\{P^{(n)}_\delta(\cdot \mid \tau = n) : \delta \in \Delta\}$ relative to S_n. That is, for each $\delta_0 < \delta_1$ with $\delta_0, \delta_1 \in \Delta$, $p^{(n)}_{\delta_1}(y^n \mid \tau = n)/p^{(n)}_{\delta_0}(y^n \mid \tau = n)$ is an increasing function of $s_n(y^n)$, on the set of all y^n with $p^{(n)}_\delta(y^{(n)} \mid \tau = n) > 0$ for some $\delta \in \Delta$.*
2. *As a consequence, for all $a > 0$,*

$$P_\delta\left(\frac{p^{(n)}_{\delta^+}(Y^n \mid \tau = n)}{p^{(n)}_{\delta^-}(Y^n \mid \tau = n)} \geq a \mid \tau = n\right) \tag{5.B.3}$$

   is increasing in δ for all a.

Proof. For the first part, note that for each y^n as above, we have:

$$\frac{p^{(n)}_{\delta_1}(y^n \mid \tau = n)}{p^{(n)}_{\delta_0}(y^n \mid \tau = n)} = \frac{p^{(n)}_{\delta_1}(y^n)}{p^{(n)}_{\delta_0}(y^n)} \cdot \frac{P^{(n)}_{\delta_1}(\tau = n \mid y^n)}{P^{(n)}_{\delta_0}(\tau = n \mid y^n)} \cdot \frac{P_{\delta_0}(\tau = n)}{P_{\delta_1}(\tau = n)} = \frac{p^{(n)}_{\delta_1}(y^n)}{p^{(n)}_{\delta_0}(y^n)} \cdot \frac{P_{\delta_0}(\tau = n)}{P_{\delta_1}(\tau = n)},$$

where the first equality is Bayes' theorem and the second equality follows, because for the type of stopping rule we employ, conditioned on the sequence y^n, the probability of

stopping exactly after having seen outcomes is independent of δ. But the rightmost expression shows that the likelihood ratio for the densities conditioned on $\tau = n$ must be an increasing function of $s_n(y^n)$ since, by assumption, the original, unconditional likelihood ratio is as well.

The second part follows immediately from the well-known connection (Lehmann, 1986) between monotone likelihood ratios and stochastic dominance; see
`https://math.stackexchange.com/questions/733291/why-mlr-monotone-likelihood-ratio-implies-stochastic-increasing` for a very short, simple, yet correct proof. □

In the remainder of the proof, we write $p_\delta(\cdot \mid \tau = n)$ instead of $p^{(n)}(\cdot \mid \tau = n)$ for brevity. Let $\mathscr{E}_{r,n}$ be the event that $\pi^{\text{ps}}(H_1 \mid Y^n)/\pi^{\text{ps}}(H_0 \mid Y^n) \geq r$. Since by the irrelevance of the stopping rule we have

$$\frac{\pi^{\text{ps}}(H_1 \mid Y^n)}{\pi^{\text{ps}}(H_0 \mid Y^n)} = \frac{\pi^{\text{ps}}(H_1 \mid \tau = n, Y^n)}{\pi^{\text{ps}}(H_0 \mid \tau = n, Y^n)} = \frac{\pi^{\text{ps}}(H_1 \mid \tau = n)}{\pi^{\text{ps}}(H_0 \mid \tau = n)} \cdot \frac{p_{\delta^+}(Y^n \mid \tau = n)}{p_{\delta^-}(Y^n \mid \tau = n)}$$

we have that $\mathscr{E}_{r,n}$ is equivalent to the event that $\frac{p_{\delta^+}(Y^n|\tau=n)}{p_{\delta^-}(Y^n|\tau=n)} \geq r \frac{\pi^{\text{ps}}(H_0|\tau=n)}{\pi^{\text{ps}}(H_1|\tau=n)}$. We then have:

$$\begin{aligned}
&\frac{\pi\left(H_1 \mid \mathscr{E}_{r,n}, \tau = n\right)}{\pi\left(H_0 \mid \mathscr{E}_{r,n}, \tau = n\right)} \overset{(a)}{=} \frac{\bar{P}_1\left(\mathscr{E}_{r,n} \mid \tau = n\right)\pi(H_1 \mid \tau = n)}{\bar{P}_0\left(\mathscr{E}_{r,n} \mid \tau = n\right)\pi(H_0 \mid \tau = n)} \overset{(b)}{\geq} \frac{P_{\delta^+}\left(\mathscr{E}_{r,n} \mid \tau = n\right)\pi(H_1|\tau = n)}{P_{\delta^-}\left(\mathscr{E}_{r,n} \mid \tau = n\right)\pi(H_0|\tau = n)} \\
&\overset{(c)}{=} \frac{\pi(H_1 \mid \tau = n) \cdot \sum_{y^n} p_{\delta^+}(y^n \mid \tau = n) \cdot \mathbf{1}_{\frac{p_{\delta^+}(y^n|\tau=n)}{p_{\delta^-}(y^n|\tau=n)} \geq \frac{\pi^{\text{ps}}(H_0|\tau=n)}{\pi^{\text{ps}}(H_1|\tau=n)} \cdot r}}{\pi(H_0 \mid \tau = n) \cdot \sum_{y^n} p_{\delta^-}(y^n \mid \tau = n) \cdot \mathbf{1}_{\frac{p_{\delta^+}(y^n|\tau=n)}{p_{\delta^-}(y^n|\tau=n)} \geq \frac{\pi^{\text{ps}}(H_0|\tau=n)}{\pi^{\text{ps}}(H_1|\tau=n)} \cdot r}} \overset{(d)}{\geq} \\
&\frac{\pi(H_1 \mid \tau = n) \cdot \left(r \cdot \frac{\pi^{\text{ps}}(H_0|\tau=n)}{\pi^{\text{ps}}(H_1|\tau=n)}\right) \cdot \sum_{y^n} p_{\delta^-}(y^n \mid \tau = n) \cdot \mathbf{1}_{\mathscr{E}_{r,n}}}{\pi(H_0 \mid \tau = n) \cdot \sum_{y^n} p_{\delta^-}(y^n \mid \tau = n) \cdot \mathbf{1}_{\mathscr{E}_{r,n}}} \overset{(e)}{=} r \cdot \frac{\bar{P}_1(\tau = n)}{\bar{P}_0(\tau = n)} \cdot \frac{P_{\delta^-}(\tau = n)}{P_{\delta^+}(\tau = n)}.
\end{aligned}$$

Here (a) is an instance of (5.A.1). We note that this inequality still holds in our generalized set-up as long as the probability of the set $\mathscr{E}_1$ under $P_{\delta,\gamma}$ does not depend on γ. (b) follows by first applying Lemma 5.B.2. (5.B.3) gives, using that $\bar{P}_1$ is a mixture of P_δ with $\delta \geq \delta^+$, that $\bar{P}_1(p_{\delta^+}(y^n)/p_{\delta^-}(y^n) \geq r \mid \tau = n) \geq P_{\delta^+}(p_{\delta^+}(y^n)/p_{\delta^-}(y^n) > r \mid \tau = n)$. Similarly it gives that $\bar{P}_0(p_{\delta^+}(y^n)/p_{\delta^-}(y^n) \geq r \mid \tau = n) \leq P_{\delta^-}(p_{\delta^+}(y^n)/p_{\delta^-}(y^n) > r \mid \tau = n)$, and then (b) follows. (c) is merely writing out the definition, (d) follows by applying the inequality in the event in the indicator function and for (e) we used Bayes' theorem again.

This chain of inequalities gives (5.B.2), thus finishing the proof for the discrete case.

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

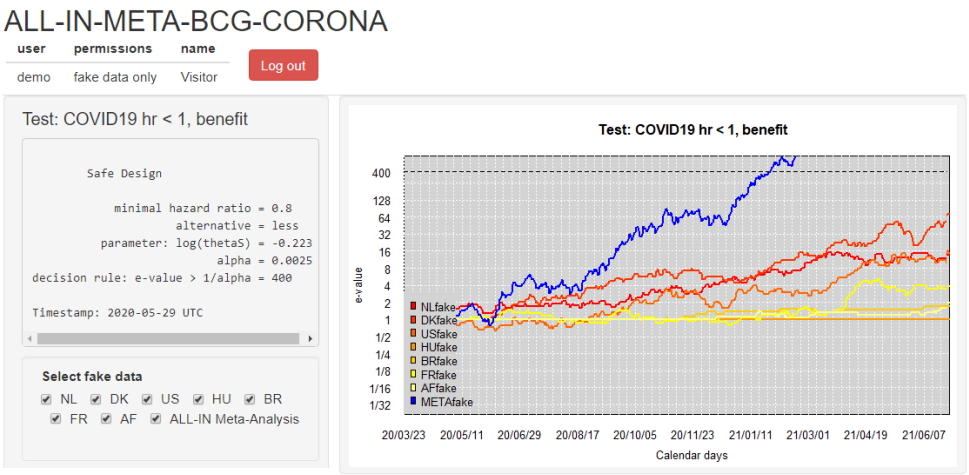

Figure 6.1. *Dashboard used to communicate interim results in ALL-IN-META-BCG-CORONA to all data uploaders with a login. The involved trials were performed in the Netherlands (NL), Denmark (DK), the United States (US), Hungary (HU), Brazil (BR), France (FR), and Guinea-Bissau/Mozambique (AF). The dashboard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

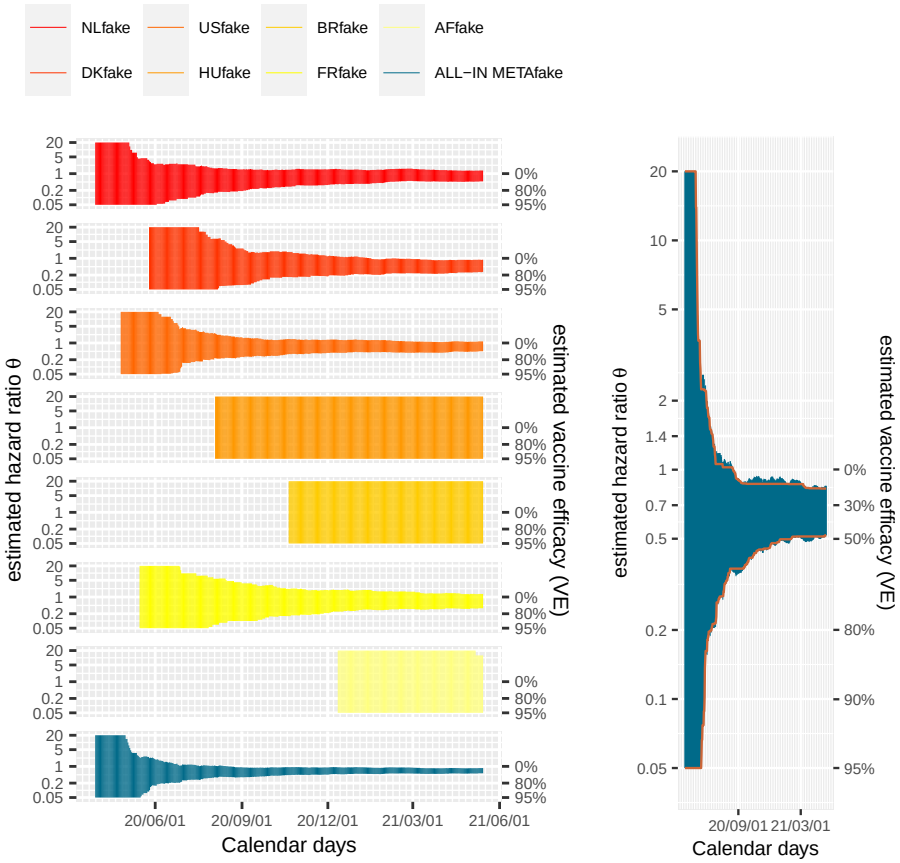

Figure 6.3. Anytime-valid confidence sequences corresponding to the fake data in Figure 6.1 with the running intersection for the meta-analysis sequence. Data generated according to the meta-analysis design with effect just slightly larger (hazard ratio = 0.7) than that of minimal interest (hazard ratio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

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

Bibliography

Akl, E. A., Meerpohl, J. J., Elliott, J., Kahale, L. A., Schünemann, H. J., Agoritsas, T., Hilton, J., Perron, C., Akl, E., Hodder, R., et al. (2017). Living systematic reviews: 4. living guideline recommendations. *Journal of clinical epidemiology*, 91:47–53.

Altman, D. G. (1994). The scandal of poor medical research. *BMJ*, 308(6924):283–284. Publisher: British Medical Journal Publishing Group Section: Editorial.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer Series in Statistics. Springer-Verlag, New York.

Armitage, P. (1984). Controversies and achievements in clinical trials. *Contemporary Clinical Trials*, 5(1):67–72.

Baum, M. L., Anish, D. S., Chalmers, T. C., Sacks, H. S., Smith Jr, H., and Fagerstrom, R. M. (1981). A survey of clinical trials of antibiotic prophylaxis in colon surgery: evidence against further use of no-treatment controls. *New England Journal of Medicine*, 305(14):795–799.

Berger, J. O. and Berry, D. A. (1988). The relevance of stopping rules in statistical inference. *Statistical decision theory and related topics IV*, 1:29–47.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., and Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. John Wiley & Sons, Ltd. DOI: 10.1002/9780470743386.refs.

Borm, G. F. and Donders, A. R. T. (2009). Updating meta-analyses leads to larger type I errors than publication bias. *Journal of clinical epidemiology*, 62(8):825–830.

Branswell, H. (2021). 12 lessons Covid-19 taught us about developing vaccines during a pandemic. https://www.statnews.com/2021/06/30/12-lessons-covid-19-developing-vaccines/. Accessed: 12 July 2021.

Breiman, L. (1961). Optimal gambling systems for favorable games. *Fourth Berkeley Symposium*.

Brok, J., Thorlund, K., Wetterslev, J., and Gluud, C. (2008). Apparently conclusive meta-analyses may be inconclusive—trial sequential analysis adjustment of random error risk due to repetitive testing of accumulating data in apparently conclusive neonatal meta-analyses. *International journal of epidemiology*, 38(1):287–298.

Chalmers, I., Bracken, M. B., Djulbegovic, B., Garattini, S., Grant, J., Gülmezoglu, A. M., Howells, D. W., Ioannidis, J. P., and Oliver, S. (2014). How to increase value and reduce waste when research priorities are set. *The Lancet*, 383(9912):156–165.

Chalmers, I. and Glasziou, P. (2009). Avoidable waste in the production and reporting of research evidence. *The Lancet*, 114(6):1341–1345.

Chalmers, I. and Glasziou, P. (2016). Systematic reviews and research waste. *The Lancet*, 387(10014):122–123.

Chalmers, T. C. and Lau, J. (1993). Meta-analytic stimulus for changes in clinical trials. *Statistical Methods in Medical Research*, 2(2):161–172.

Claxton, K. P., Sculpher, M., and Drummond, M. (2002). A rational framework for decision making by the national institute for clinical excellence (NICE). *The Lancet*, 360(9334):711–715.

Claxton, K. P. and Sculpher, M. J. (2006). Using value of information analysis to prioritise health research. *Pharmacoeconomics*, 24(11):1055–1068.

Cochrane Scientific Committee, Schmid, C., Senn, S., Sterne, J., Kulinskaya, E., Posch, M., Roes, K., and McKenzie, J. (2018). Should Cochrane apply error-adjustment methods when conducting repeated meta-analyses?

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society Series B*, 34(2):187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

Crane, H. and Shafer, G. (2020). Risk is random: The magic of the d'alembert. Technical report, Working Paper 57, Available from: https://www. proba bilit yandf inance. com.

CureVac AG (2020). Clinical trial protocol a phase 2b/3, randomized, observer-blinded, placebo-controlled, multicenter clinical study evaluating the efficacy and safety of investigational sars-cov-2 mrna vaccine cvncov in adults 18 years of age and older. https://www.curevac.com/wp-content/uploads/2021/06/HERALD_CV-NCOV-004-Protocol.pdf. Accessed: 16 July 2021.

CureVac AG (2021). Curevac final data from phase 2b/3 trial of first-generation Covid-19 vaccine candidate, cvncov, demonstrates protection in age group of 18 to 60. https://www.curevac.com/en/2021/06/30/curevac-final-data-from\-phase-2b-3-trial-of-first-generation-covid-19-vaccine\-candidate-cvncov-demonstrates-protection-in-age-group-of-18-to-60/. Accessed: 16 July 2021.

Darling, D. and Robbins, H. (1967). Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences of the United States of America*, 58(1):66.

Darling, D. and Robbins, H. (1968). Some nonparametric sequential tests with power

one. *Proceedings of the National Academy of Sciences of the United States of America*, 61(3):804.

Davey, J., Turner, R. M., Clarke, M. J., and Higgins, J. P. (2011). Characteristics of meta-analyses and their component studies in the Cochrane database of systematic reviews: a cross-sectional, descriptive analysis. *BMC medical research methodology*, 11(1):160.

Dawid, A. P. (1984). Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, 147(2):278–290.

Dawid, A. P. (1994). Selection paradoxes of Bayesian inference. *Lecture Notes-Monograph Series*, pages 211–220.

van Dongen, N. and van Grootel, L. (2021). Overview on the null hypothesis significance test. https://psyarxiv.com/hwk4n/download?format=pdf.

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek, B. A., and Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, 112(50):15343–15347.

Edwards, A. F. (1974). *Likelihood: An account of the statistical concept of likelihood and its application to scientific inference*. Cambridge University Press, New York.

Efron, B. (1974). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72(359):557–565. with discussion.

Egger, M. and Smith, G. D. (1998). Bias in location and selection of studies. *BMJ: British Medical Journal*, 316(7124):61.

Elliott, J. H., Synnot, A., Turner, T., Simmonds, M., Akl, E. A., McDonald, S., Salanti, G., Meerpohl, J., MacLehose, H., Hilton, J., et al. (2017). Living systematic review: 1. introduction—the why, what, when, and how. *Journal of clinical epidemiology*, 91:23–30.

Ellis, S. P. and Stewart, J. W. (2009). Temporal dependence and bias in meta-analysis. *Communications in Statistics—Theory and Methods*, 38(15):2453–2462.

Evans, I., Thornton, H., Chalmers, I., and Glasziou, P. (2011). *Testing treatments: better research for better healthcare*. Pinter & Martin Publishers.

Evbres (2021). Hans lund - EBR - placing research in the context of existing knowledge. https://www.youtube.com/watch?v=dko-35vvJFk&list=PLkIyhRK9IXIOzuqBwNMFl1gyyE4vvdi9u&index=3.

FDA (2020). Development and Licensure of Vaccines to Prevent Covid-19. https://www.fda.gov/media/139638/download. Accessed: 12 July 2021.

FDA (2021). Emergency Use Authorization for Vaccines to Prevent Covid-19. https://www.fda.gov/media/142749/download. Accessed: 12 July 2021.

Fergusson, D., Glass, K. C., Hutton, B., and Shapiro, S. (2005). Randomized controlled trials of aprotinin in cardiac surgery: could clinical equipoise have stopped the bleeding? *Clinical Trials*, 2(3):218–232.

Fisher, R. A. (1938). Presidential address. *Sankhyā: The Indian Journal of Statistics*, pages 14–17.

Gehr, B. T., Weiss, C., and Porzsolt, F. (2006). The fading of reported effectiveness. a meta-analysis of randomised controlled trials. *BMC medical research methodology*, 6(1):25.

Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science*, 1(2):198–218.

Glasziou, P. and Chalmers, I. (2018). Research waste is still a scandal—an essay by Paul Glasziou and Iain Chalmers. *BMJ*, 363. Publisher: British Medical Journal Publishing Group Section: Feature.

Glasziou, P., Sanders, S., and Hoffmann, T. (2020). Waste in Covid-19 research. *BMJ*, 369. Publisher: British Medical Journal Publishing Group Section: Editorial.

Gøtzsche, P. C. (1987). Reference bias in reports of drug trials. *Br Med J (Clin Res Ed)*, 295(6599):654–656.

Goudie, A. C., Sutton, A. J., Jones, D. R., and Donald, A. (2010). Empirical assessment suggests that existing evidence could be used more fully in designing randomized controlled trials. *Journal of Clinical Epidemiology*, 63:983–991.

Grünwald, P. (2021). Peter D. Grünwald's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(2):440–441.

Grünwald, P., de Heide, R., and Koolen, W. (2019). Safe testing. *arXiv preprint arXiv:1906.07801*.

Grünwald, P. and Mehta, N. (2019). A tight excess risk bound via a unified PAC-Bayesian-Rademacher-Shtarkov-MDL complexity. In *Proceedings of the Thirtieth Conference on Algorithmic Learning Theory (ALT) 2019*.

Grünwald, P. and Roos, T. (2020). Minimum Description Length revisited. *International Journal of Mathematics for Industry*, 11(1).

de Heide, R. and Grünwald, P. D. (2021). Why optional stopping can be a problem for Bayesians. *Psychonomic Bulletin & Review*, 28(3):795–812.

Hendriksen, A., de Heide, R., and Grünwald, P. (2020). Optional stopping with Bayes factors: a categorization and extension of folklore results, with an application to invariant situations. *Bayesian Analysis*.

Henzi, A. and Ziegel, J. F. (2021). Valid sequential inference on probability forecast performance. *arXiv preprint arXiv:2103.08402*.

Higgins, J. P., Whitehead, A., and Simmonds, M. (2011). Sequential methods for random-effects meta-analysis. *Statistics in medicine*, 30(9):903–921.

Howard, S. R. and Ramdas, A. (2019). Sequential estimation of quantiles with applications to A/B-testing and best-arm identification. *arXiv preprint arXiv:1906.09712*.

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2018). Exponential line-crossing inequalities. *arXiv:1808.03204 [math]*. arXiv: 1808.03204.

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080.

Imberger, G., Thorlund, K., Gluud, C., and Wetterslev, J. (2016). False-positive findings in cochrane meta-analyses with and without application of trial sequential analysis: an empirical review. *BMJ open*, 6(8):e011890.

Ioannidis, J. P. (2005a). Contradicted and initially stronger effects in highly cited clinical research. *Jama*, 294(2):218–228.

Ioannidis, J. P. (2005b). Why most published research findings are false. *PLoS medicine*, 2(8):e124.

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology*, pages 640–648.

Ioannidis, J. P. (2010). Meta-research: The art of getting it wrong. *Research Synthesis Methods*, 1(3-4):169–184.

Ioannidis, J. P. and Trikalinos, T. A. (2005). Early extreme contradictory estimates may appear in published research: the Proteus phenomenon in molecular genetics research and randomized trials. *Journal of clinical epidemiology*, 58(6):543–549.

Jackson, D. and Turner, R. (2017). Power analysis for random-effects meta-analysis. *Research synthesis methods*, 8(3):290–302.

Johari, R., Koomen, P., Pekelis, L., and Walsh, D. (2021). Always valid inference: Continuous monitoring of A/B tests. *Operations Research*.

Kelly, J. (1956). A new interpretation of information rate. *Bell System Technical Journal*, pages 917–926.

Klein, J. P. and Moeschberger, M. L. (2006). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer Science & Business Media. Google-Books-ID: aO7xBwAAQBAJ.

Konnikova, M. (2020). *The Biggest Bluff: How I Learned to Pay Attention, Master Myself, and Win*. Penguin.

Krum, H. and Tonkin, A. (2003). Why do phase III trials of promising heart failure drugs often fail? the contribution of "regression to the truth". *Journal of cardiac failure*, 9(5):364–367.

Kulinskaya, E., Huggins, R., and Dogo, S. H. (2016). Sequential biases in accumulating evidence. *Research synthesis methods*, 7(3):294–305.

Kulinskaya, E. and Wood, J. (2014). Trial sequential methods for meta-analysis. *Research synthesis methods*, 5(3):212–220.

Lai, T. L. (1976). On confidence sequences. *The Annals of Statistics*, 4(2):265–280.

Lakens, D. (2021). Sample size justification. https://psyarxiv.com/9d3yf/download?format=pdf.

Lan, K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, 70(3):659–663.

Lau, J., Antman, E. M., Jimenez-Silva, J., Kupelnick, B., Mosteller, F., and Chalmers, T. C. (1992). Cumulative meta-analysis of therapeutic trials for myocardial infarction. *New England Journal of Medicine*, 327(4):248–254.

Lau, J., Schmid, C. H., and Chalmers, T. C. (1995). Cumulative meta-analysis of clinical trials builds evidence for exemplary medical care. *Journal of clinical epidemiology*, 48(1):45–57.

Lawrence, J. M., Meyerowitz-Katz, G., Heathers, J. A., Brown, N. J., and Sheldrick, K. A. (2021). The lesson of ivermectin: meta-analyses based on summary data alone are inherently unreliable. *Nature Medicine*, pages 1–2.

Lehmann, E. (1986). *Testing statistical hypotheses*. Wiley.

Li, J. and Barron, A. (2000). Mixture density estimation. In Solla, S., Leen, T., and Müller, K.-R., editors, *Advances in Neural Information Processing Systems*, volume 12, pages 279–285, Cambridge, MA. MIT Press.

Li, Q. J. (1999). *Estimation of Mixture Models*. PhD Thesis, Yale University, New Haven, CT, USA.

Lund, H., Brunnhuber, K., Juhl, C., Robinson, K., Leenaars, M., Dorch, B. F., Jamtvedt, G., Nortvedt, M. W., Christensen, R., and Chalmers, I. (2016). Towards evidence based research. *Bmj*, 355:i5440.

Mallett, S. and Clarke, M. (2003). How many Cochrane reviews are needed to cover existing evidence on the effects of health care interventions? *ACP journal club*, 139(1):A11–A11.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother. Rep.*, 50:163–170.

McDonald, A. M., Knight, R. C., Campbell, M. K., Entwistle, V. A., Grant, A. M., Cook, J. A., Elbourne, D. R., Francis, D., Garcia, J., Roberts, I., et al. (2006). What influences recruitment to randomised controlled trials? a review of trials funded by two UK funding agencies. *Trials*, 7(1):1–8.

McShane, B. B. and Gal, D. (2017). Statistical significance and the dichotomization of evidence. *Journal of the American Statistical Association*, 112(519):885–895.

Mehrotra, D. V. and Roth, A. J. (2001). Relative risk estimation and inference using a generalized logrank statistic. *Statistics in Medicine*, 20(14):2099–2113. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.854.

Moher, D., Tetzlaff, J., Tricco, A. C., Sampson, M., and Altman, D. G. (2007a). Epidemiology and reporting characteristics of systematic reviews. *PLoS medicine*, 4(3):e78.

Moher, D. and Tsertsvadze, A. (2006). Systematic reviews: when is an update an update? *The Lancet*, 367(9514):881–883.

Moher, D., Tsertsvadze, A., Tricco, A., Eccles, M., Grimshaw, J., Sampson, M., and Barrowman, N. (2008). When and how to update systematic reviews. *Cochrane database of systematic reviews*, (1).

Moher, D., Tsertsvadze, A., Tricco, A. C., Eccles, M., Grimshaw, J., Sampson, M., and Barrowman, N. (2007b). A systematic review identified few methods and strategies describing when and how to update systematic reviews. *Journal of clinical epidemiology*, 60(11):1095–e1.

Netea, M. G., Giamarellos-Bourboulis, E. J., Domínguez-Andrés, J., Curtis, N., van Crevel, R., van de Veerdonk, F. L., and Bonten, M. (2020). Trained immunity: a tool for reducing susceptibility to and the severity of sars-cov-2 infection. *Cell*, 181(5):969–977.

Newman, M. (2019). Has cochrane lost its way? *Bmj*, 364.

O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, pages 549–556.

Pace, L. and Salvan, A. (2019). Likelihood, replicability and Robbins' confidence sequences. *International Statistical Review*.

Page, M. J., Shamseer, L., Altman, D. G., Tetzlaff, J., Sampson, M., Tricco, A. C., Catalá-López, F., Li, L., Reid, E. K., Sarkis-Onofre, R., et al. (2016). Epidemiology and reporting characteristics of systematic reviews of biomedical research: a cross-sectional study. *PLoS medicine*, 13(5):e1002028.

Pereira, T. V. and Ioannidis, J. P. (2011). Statistically significant meta-analyses of clinical trials have modest credibility and inflated effects. *Journal of clinical epidemiology*, 64(10):1060–1069.

Peto, R. (1972). Discussion on the paper 'Regression models and Life Tables by Sir David R. Cox. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 34(2):205–208.

Peto, R. (1987). Why do we need systematic overviews of randomized trials?(transcript of an oral presentation, modified by the editors). *Statistics in medicine*, 6(3):233–240.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society: Series A (General)*, 135(2):185–198.

Pfeiffer, T., Bertram, L., and Ioannidis, J. P. (2011). Quantifying selective reporting and the Proteus phenomenon for multiple datasets with similar bias. *PLoS One*, 6(3):e18362.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199.

Pocock, S. J. (2006). Current controversies in data monitoring for clinical trials. *Clinical trials*, 3(6):513–521.

Polack, F. P., Thomas, S. J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., Perez, J. L., Marc, G. P., Moreira, E. D., Zerbini, C., et al. (2020). Safety and efficacy of the NT162b2 mRNA Covid-19 vaccine. *New England Journal of Medicine*.

Polanin, J. R. and Williams, R. T. (2016). Overcoming obstacles in obtaining individual participant data for meta-analysis. *Research synthesis methods*, 7(3):333–341.

Potthoff, R. F. (2007). Prediction markets, Bayesian priors, and clinical trials. *Journal of statistical planning and inference*, 137(11):3706–3721.

Proschan, M. A., Lan, K. G., and Wittes, J. T. (2006). *Statistical monitoring of clinical trials: a unified approach*. Springer Science & Business Media.

Ramdas, A., Ruf, J., Larsson, M., and Koolen, W. (2020). Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv preprint arXiv:2009.03167*.

Riley, R. D., Higgins, J. P., and Deeks, J. J. (2011). Interpretation of random effects meta-analyses. *Bmj*, 342.

Robbins, H. (1970). Statistical methods related to the law of the iterated logarithm. *Annals of Mathematical Statistics*, 41:1397–1409.

Roberts, I. and Ker, K. (2015). How systematic reviews cause research waste. *The Lancet*, 386(10003):1536.

Robinson, K. A. and Goodman, S. N. (2011). A systematic examination of the citation of prior research in reports of randomized, controlled trials. *Annals of internal medicine*, 154(1):50–55.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin*, 86(3):638.

Rothman, K. J., Gallacher, J. E., and Hatch, E. E. (2013). Why representativeness should be avoided. *International journal of epidemiology*, 42(4):1012–1014.

Royall, R. (1997). *Statistical evidence: a likelihood paradigm*, volume 71. Chapman & Hall/CRC.

Royall, R. (2000). On the probability of observing misleading statistical evidence. *Journal of the American Statistical Association*, 95(451):760–768.

Schoenfeld, D. (1981). The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika*, 68(1):316–319.

ter Schure, J. (2019). Code for paper Accumulation bias in meta-analysis: The need to consider time in error control. https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:127617.

ter Schure, J. (2021a). Code for blogposts Accumulation Bias: How to handle it. https://osf.io/p2rtw/.

ter Schure, J. (2021b). Code for paper ALL-IN meta-analysis: breathing life into living systematic reviews. https://osf.io/d9jny/.

ter Schure, J. (2021c). Judith ter Schure's contribution to the Discussion of 'Testing by betting: A strategy for statistical and scientific communication' by Glenn Shafer. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(2):440–441.

ter Schure, J. and Grünwald, P. (2019). Accumulation Bias in meta-analysis: the need to consider time in error control [version 1; peer review: 2 approved]. *F1000Research*, 8:962.

ter Schure, J., Ly, A., Pérez-Ortiz, M. F., and Grünwald, P. (2020a). Safestats and ALL-IN meta-analysis project page. https://projects.cwi.nl/safestats/.

ter Schure, J. and Pérez-Ortiz, M. F. (2022). Code for paper Safe logrank test. https://osf.io/3n8g2/.

ter Schure, J., Pérez-Ortiz, M. F., Ly, A., and Grünwald, P. (2020b). The safe logrank test: Error control under continuous monitoring with unlimited horizon. *arXiv preprint arXiv:2011.06931*.

Sellke, T. and Siegmund, D. (1983). Sequential analysis of the proportional hazards model. *Biometrika*, 70(2):315–326.

Senn, S. (2000). The many modes of meta. *Drug Information Journal*, 34(2):535–549.

Senn, S. (2008). A note concerning a selection "paradox" of Dawid's. *The American Statistician*, 62(3):206–210.

Senn, S. (2014). A note regarding meta-analysis of sequential trials with stopping for efficacy. *Pharmaceutical Statistics*, 13(6):371–375.

Shafer, G. (2019). The language of betting as a strategy for statistical and scientific communication. http://probabilityandfinance.com/articles/54.pdf. Accessed 16 May 2019.

Shafer, G. (2021). Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(2):407–431.

Shafer, G., Shen, A., Vereshchagin, N., and Vovk, V. (2011). Test martingales, Bayes factors and p-values. *Statistical Science*, 26(1):84–101.

Shafer, G. and Vovk, V. (2019). *Game-Theoretic Foundations for Probability and Finance*, volume 455. John Wiley & Sons.

Shamy, M., Dewar, B., and Fedyk, M. (2020). Different meanings of equipoise and the four quadrants of uncertainty. *Journal of Clinical Epidemiology*, 127:248–249.

Simmonds, M., Salanti, G., McKenzie, J., and Elliott, J. (2017). Living systematic reviews: 3. statistical methods for updating meta-analyses. *Journal of clinical epidemiology*, 91:38–46.

Simmonds, M. C., Tierney, J., Bowden, J., and Higgins, J. P. (2011). Meta-analysis of time-to-event data: a comparison of two-stage methods. *Research synthesis methods*, 2(3):139–149.

SIPTA, L. (2021). Game-theoretic foundations for statistical testing: Glenn shafer. https://www.youtube.com/watch?v=pOMTb_x2mxw.

Slud, E. V. (1984). Sequential linear rank tests for two-sample censored survival data. *The Annals of Statistics*, pages 551–571.

Slud, E. V. (1992). Partial likelihood for continuous-time stochastic processes. *Scandinavian journal of statistics*, pages 97–109.

Sutton, A. J., Cooper, N. J., Jones, D. R., Lambert, P. C., Thompson, J. R., and Abrams, K. R. (2007). Evidence-based sample size calculations based upon updated meta-analysis. *Statistics in Medicine*, 26(12):2479–2500. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.2704.

Tierney, J. F., Fisher, D. J., Vale, C. L., Burdett, S., Rydzewska, L. H., Rogozińska, E., Godolphin, P. J., White, I. R., and Parmar, M. K. (2021). A framework for prospective, adaptive meta-analysis (FAME) of aggregate data from randomised trials. *PLoS medicine*, 18(5):e1003629.

Turner, R., Ly, A., and Grünwald, P. (2021). Safe tests and always-valid confidence intervals for contingency tables and beyond. *arXiv preprint arXiv:2106.02693*.

Turner, R., Ly, A., Pérez-Ortiz, M. F., ter Schure, J., and Grünwald, P. (2022). *R-package* `safestats`. R package version 0.8.6, Maintainer: Alexander Ly <a.ly@jaspstats.org>, https://cran.r-project.org/package=safestats.

Turner, R. M., Bird, S. M., and Higgins, J. P. (2013). The impact of study size on meta-analyses: examination of underpowered studies in Cochrane reviews. *PloS one*, 8(3):e59202.

Ville, J. (1939). *Etude critique de la notion de collectif*. Gauthier-Villars, Paris.

Vovk, V. and Wang, R. (2021). E-values: Calibration, combination, and applications. *Annals of Statistics*.

Wald, A. (1947). Sequential analysis. 1947. *Zbl0029*, 15805.

Walters, S. J., dos Anjos Henriques-Cadby, I. B., Bortolami, O., Flight, L., Hind, D., Jacques, R. M., Knox, C., Nadin, B., Rothwell, J., Surtees, M., et al. (2017). Recruitment and retention of participants in randomised controlled trials: a review of trials funded and

published by the united kingdom health technology assessment programme. *BMJ open*, 7(3):e015276.

Wang, M. Q., Yan, A. F., and Katz, R. V. (2018). Researcher requests for inappropriate analysis and reporting: A US survey of consulting biostatisticians. *Annals of internal medicine*, 169(8):554–558.

van Werkhoven, C. H., ter Schure, J., Bonten, M., Netea, M., Grünwald, P., and Ly, A. (2021). Anytime Live and Leading Interim meta-analysis of the impact of Bacillus Calmette-Guérin vaccination in health care workers and elderly during the sars-cov-2 pandemic (ALL-IN-META-BCG-CORONA). https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=213069&VersionID=1473878.

Wetterslev, J., Jakobsen, J. C., and Gluud, C. (2017). Trial sequential analysis in systematic reviews with meta-analysis. *BMC medical research methodology*, 17(1):39.

Wetterslev, J., Thorlund, K., Brok, J., and Gluud, C. (2008). Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. *Journal of clinical epidemiology*, 61(1):64–75.

Whitehead, A. (1997). A prospectively planned cumulative meta-analysis applied to a series of concurrent clinical trials. *Statistics in medicine*, 16(24):2901–2913.

Whitehead, A. (2002). *Meta-analysis of controlled clinical trials*, volume 7. John Wiley & Sons.

van der Wier, M. (2020). Proef: Helpt tbc-vaccin in de strijd tegen het coronavirus? https://www.trouw.nl/nieuws/proef-helpt-tbc-vaccin-in-de-strijd\-tegen-het-coronavirus~b2f670df/.

Wu, J. and Xiong, X. (2017). Group sequential survival trial design and monitoring using the log-rank test. *Statistics in biopharmaceutical research*, 9(1):35–43.

Young, C. and Horton, R. (2005). Putting clinical trials into context. *The Lancet*, 366(9480):107–108. Publisher: Elsevier.

Yu, L.-M., Bafadhel, M., Dorward, J., Hayward, G., Saville, B. R., Gbinigie, O., Van Hecke, O., Ogburn, E., Evans, P. H., Thomas, N. P., et al. (2021). Inhaled budesonide for Covid-19 in people at high risk of complications in the community in the UK (PRINCIPLE): a randomised, controlled, open-label, adaptive platform trial. *The Lancet*, 398(10303):843–855.

Yusuf, S., Peto, R., Lewis, J., Collins, R., and Sleight, P. (1985). Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Progress in Cardiovascular Diseases*, 27(5):335–371.


20
50
EURO

Samenvatting

Wetenschappelijke kennis vermeerdert zich veel te inefficiënt. Het is vaak een lappendeken van onderzoeksbijdragen zonder gezamenlijke afstemming. Met name in klinische trials (zogenaamde RCT's) zijn de vervolgstudies die worden gedaan niet altijd het meest beloftevol. Ook zijn ze niet altijd ontworpen voor de extra bewijslast die nodig is. Is dat wel het geval, dan maakt standaard statistiek het onmogelijk om met die slimmigheden rekening te houden.

Dit proefschrift gaat over het stapelen van wetenschappelijke inzichten, over de statistische problemen met standaard methodes (accumulation bias) en over nieuwe statistische methodes om het beter te doen. We kunnen resultaten efficiënt samen nemen in zogenaamde meta-analyse door ALL-IN te gaan. Wetenschap is namelijk altijd een gok: aan het begin van een nieuwe studie is er weinig zekerheid. Maar gokken kan strategisch, en voor klinische trials kunnen we eerdere resultaten gebruiken om te bepalen of de nieuwe studie noodzakelijk is en optimaal ontworpen.

In de wetenschap komt succes van studies die vragen beantwoorden; die onderscheid kunnen maken tussen data die voor een nieuw wetenschappelijk idee spreekt, en data die het tegen spreekt. Door slim te stapelen dwing je succes af om dat onderscheid te maken. Zoals professioneel pokeraars in een reeks pokertoernooien gebruik maken van een bankrekening vol reservegeld. Dat geld verzamelen ze met eerdere successen, ze vestigen zich ermee als professional en ze winnen pokertoernooien door all-in te gaan.

ALL-IN meta-analyse staat voor *Anytime*, *Live* and *Leading INterim* meta-analyse. Het kan wetenschappers helpen om onderzoek te prioriteren, nieuwe resultaten te interpreteren in de context van wat er al is, en zo efficiënt hun gok te wagen met nieuw onderzoek. De noodzaak hiervan is geïnspireerd door de replicatiecrisis in de empirische wetenschap en de beweging om Research Waste (onderzoeksverspilling) tegen te gaan in de biomedische wetenschappen. Maar er zijn meer voordelen. ALL-IN meta-analyse kan bottom-up samenwerking bevorderen. Statistische resultaten worden eenvoudiger om te communiceren. Bovendien stelt ALL-IN meta-analyse centraal dat wetenschap niet om individuele papers zou moeten gaan. Met ALL-IN meta-analyse wordt het weer een voortdurend proces van tussentijdse resultaten, noodzakelijke beslismomenten en communicatie tussen vakgenoten.

Dankwoord

Veel vrienden en familie leefden met mij mee in mijn promotietraject. Ze deden dappere pogingen om de p-waarde te begrijpen (en samen te verketteren), ze kwamen kijken bij de Nacht van de Wetenschap/Science Battle/Pint of Science, ze hielpen me met mijn artikeltjes en blogs of stuurden me enthousiaste berichtjes als ze gepubliceerd waren – ze konden zelfs mij uitleggen waarom het belangrijk was. Ik hoop dat ze er mogen zijn op 7 april 2022. Niet alleen om hen het Leidse academiegebouw te laten zien, maar om iedereen te laten zien wie zíj zijn!

Dat zou niet kunnen zonder mijn promotor, Peter Grünwald. Ik kende hem al als de professor die met een journalist durfde te praten, die had geholpen Lucia de B. uit de cel te krijgen en dat niet alleen deed vanwege het directe effect van statistiek op de samenleving, maar ook omdat er fundamentele statistische kwesties aan de hand waren. Ondertussen heeft Peter zijn blik van buitenaf altijd bewaard – zijn carrière bouwde hij in de informatietheorie en theoretische machine learning – en daarom zie ik hem als dé expert in Nederland op het gebied van de fundamenten van statistiek. Ik heb echt enorm veel van hem geleerd.

Statistiek als gokken was voor Peter niet alleen een fundamenteel idee, het ging hem juist om de handvaten voor wetenschappers: concrete statistische oplossingen, zoals een t-test gebaseerd op gokken en een test voor 2x2 tabellen. Toch liet hij mijn interesse gaan waar het heen ging, en zo kwam er dus een methode voor meta-analyse – ALL-IN – en legde hij het eerste contact om het écht te gaan toepassen in de begindagen van de coronapandemie. Ik gun iedereen zo'n promotor. Een promotor die je bovendien op congres in Zwitserland een berichtje stuurt: "Hoe ging het? Ik ben benieuwd :-) (maar alleen als je tijd hebt, als je nog aan het skiën bent gaat dat even voor!) Groeten Peter"

Mijn tweede promotor, Daniel Lakens, heeft me ook cruciale zetjes gegeven. Hij vertelde al zijn volgers op Twitter dat mijn eerste paper over accumulation bias de moeite waard was en liet me kennismaken met zijn geweldige netwerk van methoden- en open science hervormers. Hij leerde me dat wetenschap niet alleen draait om goede ideeën. Soms moet je zelf iets gedaan krijgen, zoals hij deed toen Nederland als eerste ter wereld een ronde wetenschapsfinanciering kreeg voor replicatieonderzoek. In het begin van de coronapandemie vond ik nog zo iemand: Henri van Werkhoven, die het aandurfde om samen de schouders te zetten onder een live meta-analyse; misschien wel de eerste in de geschiede-

nis. Met dank aan Marc Bonten, Mihai Netea, Lina Gurskaite en mijn collega Alexander voor hun vertrouwen en hulp.

Hoe een groot wiskundige eruit ziet weet ik dankzij mijn collega's: Alexander, Muriel, Wouter, Alice, Rianne, Rosanne, Allard en Tom, ik ben vaak enorm van jullie onder de indruk en vind het geweldig dat jullie je talent en training inzetten voor de wetenschap. Fijn is ook de strategie van het CWI, die steunt dat CWI'ers het soms net een beetje anders willen doen. Grote dank aan de ambitieuze open science collega's van de bibliotheek en de trouwe IT'ers van systeembeheer. En niet te vergeten de enthousiaste receptiemensen en betrokken collega's van ons restaurant (Rob, ik zal je naam nooit vergeten!).

Dankzij Marjolein, Kilian, Rinske, Thomas, Elsa en Arnold heeft dit boekje zo'n mooie foto op de kaft. Creatieve ideeën zijn er altijd in Steenwijkerwold, waar ik ben opgegroeid. Ik ben Irene en Arnold, mijn ouders, en Anneke, Elsa en Rinske, mijn zusjes, enorm dankbaar dat ze alle gekkigheid aanmoedigen. Ook mijn nichtje Noor (2 jaar) helpt me herinneren om grenzen te verkennen: maximaal met de roltrap voordat je een winkel uit moet, head-first van de glijbaan, of voor het eerst logeren bij oma samen met tante Judith.

Over gekkigheid gesproken: Laura en Martijn, jullie zijn knetter. Knetter dat jullie mijn promotieonderzoek zo mateloos interessant vinden, dat je komt met het woord "nemesis" voor in het voorwoord en dat ik jullie mag bedanken omdat "zonder ons had je niet kunnen zeggen dat er zonder jou geen proefschrift was geweest". Het is geweldig dat jullie er zo voor me zijn!

Marnick, best indrukwekkend dat je mijn onderzoek kunt uitleggen met al die halfbakken ideeën die je hebt aangehoord! Het helpt zoveel om die hardop te zeggen. (Niet alles blijft even goed hangen, zoals het concept supermartingaal – bij jou lang bekend als "Super Mario!" – al wilde je wel samen op vakantie naar Martigues!) Ik ben je heel erg dankbaar, niet in de minste plaats voor je kritische proeflezen en steun bij de laatste loodjes. Dankjewel dat je ook houdt van CO2-besparing/€-excelsheets, dat je me coacht om een betere klimmer te worden en dat je je skills als vakantieplanner inzet zodat ik af en toe eventjes *uit* kan.

Curriculum Vitae

Judith ter Schure's research interests lie in the foundations of statistics as well as in applied work. She has been dividing her time between her Ph.D. research on ALL-IN meta-analysis at CWI, freelance statistical consultancy (Significant Help), and board membership (treasurer) of the Netherlands Society for Statistics and Operations Research (VVSOR).

Her general motivation is the effect of statistics on society, which also inspires occasional writing for a wide audience – previously published by De Correspondent – and participation in popular science events like Nacht van de Wetenschap, Science Battle, and Pint of Science.

Judith was born in 1992 in the hospital of Meppel. During her high school Gymnasium education at **RSG Tromp Meesters in Steenwijk (SEPT 2004 - MAY 2010)**, choosing between the sciences and humanities proved too difficult. So she cobbled up a collection of subjects including mathematics, chemistry, physics, and biology, with economics and history – although she cheated the languages by choosing Latin for its dictionary and similarity to math and philosophy instead of brute force remembering words. She now has to look up how to spell "brute force". A spell checker and calculator stay must-haves to prepare any writing or calculating in public. Still, good preparation is good practice, as she learned to be a committee and board member of mountain sports associations (USAC, NSAC, NKBV) during her university studies. It is this experience that she still builds on in governing the professional society VVSOR and her homeowner's society of 32 owners and tenants; building towards ideals of professionalism, sustainable involvement of members/owners, and energy efficiency.

Judith has always been fascinated by systems of argument and decision making and wrote a high school dissertation on rhetoric ("profielwerkstuk", *cum laude*) and a Bachelor's thesis on logical fallacies. She had to admit that there are limits to the logic of language, however. Most decisions have to resort to data and do not have a written truth (to paraphrase the "In God we trust, all others bring data" W. Edwards Deming quote). This moment arrived during a **Bachelor's program in Artificial Intelligence at Utrecht University (SEPT 2010 - MARCH 2014, *cum laude*)** that combined formal logic, philosophy, linguistics, and psychology with mathematics and computer science. On a Leiden University

website, she read that according to Hal R. Varian (Google's Chief Economist, 2009) "the sexy job in the next ten years will be statisticians (sic)". She thereupon entered the **Masters program in Statistical Science at Leiden University (SEPT 2014 - MARCH 2017, *cum laude*)** and completed this Ph.D. dissertation with the Leiden Alma Mater as well **(MAY 2017 - JAN 2022)**[4].

It proved very fortunate to find her Ph.D. position with professor Peter Grünwald in his Machine Learning group at an institute that encouraged uncompromising ideals: CWI, the Netherlands institute for mathematics and computer science. Peter shared her interest in statistics' influence on society, helped her believe in the sweeping possibilities of their research results, and above all, taught her the value of the mathematical precision to produce them (and was ruthless to help her learn from her mistakes). With Peter, she got a very thorough education into the fundamentals of statistics and how to ground their new approach. Applications presented themselves in the clinical sciences, motivated by research waste opponents and the early work of evidence-based medicine enthusiasts (often the same very inspiring people). She was drawn to clinical biostatistics and specifically the regulatory clinical trials she believed to be the realm of statistical decision-making, only to find out that "statisticians need to become more like data scientists". This was uttered by EMA head of data analytics Dr. Peter Arlett, at a regulatory statistics meeting in 2021. (EMA is the European Medicines Agency that she believed to be the center of statistical decision-making to regulate new medicines.) This interesting world of data had lost its appetite for statistical decision-making, while she had just gained that appetite!

The large amounts of data going into Covid-19 pandemic decisions had solidified her love for statistics nevertheless, especially the kind that designs well-controlled randomized trials. Many interesting statistical problems remain and solving them can immensely improve scientific and medical decision making, as she learned by reading blogs and books by Stephen Senn. These lured her back to language; to the many great stories that still need to be told. Some of the best decisions during the Covid-19 pandemic were made by those designing the Covid-19 vaccine trials, and effective treatment platform trials like RECOVERY. Honestly, if statisticians need to be like anything other than statisticians, it is like the doctors and funders that pushed for well-designed, large, and simple clinical trials. A bit more ambition and initiative wouldn't hurt the profession. It can be a very sexy job indeed.

[4] Four full-time equivalent years: between May 1st, 2017 and February 1th, 2022 she spent 44 months working 80% of her working week ($\approx$ 35 weeks full-time equivalent) on this Ph.D. research and 13 months working 100%, so 48 full-time months in total.


25
50