Accumulation Bias in meta-analysis: the need to consider *time* in error control

Judith ter Schure[1] **and Peter Grünwald**[2]

[1]**CWI, Science Park 123, 1098 XG Amsterdam, The Netherlands, schure@cwi.nl (corresponding author)**
[2]**CWI, Science Park 123, 1098 XG Amsterdam, The Netherlands, pdg@cwi.nl**

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.

Keywords
meta-analysis, accumulation bias, sequential, cumulative, living systematic review, likelihood ratio, research waste, evidence-based research

1 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 paper 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 decision 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 specifically, 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 paper 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, right after we give a general introduction to Accumulation Bias in Section 2 with its relation to publication bias (Section 2.1) and an informal characterization of the direction of the bias (Section 2.2). By presenting its diversity, we argue throughout the paper 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 4) 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 5.1) and surviving over time (Section 5.2). 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 5.3). After an intermezzo on evidence that indeed such processes are already at play in Section 6, 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 7.1) as well as surviving over time (Section 7.2). The reader is left to choose between the two; the consequences of either preference are specified in Section 8. We try to give intuition on why both are still possible in their respective sections 7.1 and 7.2, but also give some extra intuition on the magic of likelihood ratios in Section 9: 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.

2 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 on series of studies of 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, and that doing so 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-analysis 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.

2.1 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 paper 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.

2.2 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 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 5.3 and 6.1. We introduce the wide variety of possible dependencies in an *Accumulation Bias Framework* in Section 4, 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 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 denote 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 and 3.3, after we simplified our meta-analysis setting to common/fixed-effects meta-analysis in Section 3.1. We present the resulting bias in the test estimates in Section 3.4 and its additional effects on the sampling distribution and testing in Section 3.5 and 3.6. In Section 3.7 we conclude by pointing out the very mild condition needed for some form of *Gold Rush* Accumulation Bias to occur

3.1 Common/fixed-effect meta-analysis

This paper 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 paper. The reason for simplification is to reduce the complexity in quantifying the problem, part of showing that quantification is not enough. In a future paper we will study the effects of heterogeneity on testing in more detail. For an ex-

ample 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 Eq. (3.1a) and in Eq. (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 A.1 for a derivation from the mean difference notation in Borenstein et al. (2009).

3.2 *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:

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

for all $t \geq 2$: (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 ($\omega_{\text{S}}^{(1)}$, $\omega_{\text{X}}^{(1)}$ and $\omega_{\text{NS}}^{(1)}$) 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 ($\omega_{\text{X}}^{(1)} = \omega_{\text{X}} = 0$) and new studies are always performed after positive significant findings, the satisfactory result ($\omega_{\text{S}}^{(1)} = \omega_{\text{S}} = 1$). Nonsignificant results have a small, but not negligible probability to spur new studies ($\omega_{\text{NS}}^{(1)} = 0.1$, $\omega_{\text{NS}} = 0.02$).

3.3 *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 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 Eq. (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 $\omega_{\text{S}}^{(1)}$ and ω_{S} given in Eq. (3.2) support the simulation of large study series under the null hypothesis. The small values for $\omega_{\text{NS}}^{(1)}$ and ω_{NS} are chosen such that the effect of significant findings on the sampling distribution is clearly visible (see Section 3.5 and Figure 1). For $\alpha = 0.05$, $\omega_{\text{S}}^{(1)} = 1$ 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 $\omega_{\text{NS}}^{(1)} = 0.1$ indicates that, since an expected 95% ($1-\alpha$) of pilot studies is not significant under the null hypothesis, 9.5% (0.1·95%) 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 $\omega_{\text{S}} = 0.02$ 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 $2.5/(2.5 + 1.9) = 0.6$.

Table 1. Expected Z-scores under the null hypothesis in the *Gold Rush* scenario, under the equal study size assumption, calculated using Eq. (3.4b) with $\alpha = 0.05$ and values for $\omega_S^{(1)}$, $\omega_{NS}^{(1)}$, ω_S and ω_{NS} from Eq. (3.2). $Z^{(t)}$ is as defined in Eq. (3.1b). See Appendix A.7 for the code that was used to calculate these values.

Number of studies (t)	$\mathbf{E}_0[Z_t]$	$\mathbf{E}_0[Z_t \mid T \geq t+1]$	$\mathbf{E}_0\left[Z^{(t)} \mid T \geq t\right]$
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

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

The new study probability parameters in Eq. (3.2) are much larger when results are positively significant than when they are not. As a result, study series that contain more significant studies have larger probabilities to come into existence than those that contain less. 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 Eq. (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 Eq. (3.2). The positive significant and nonsignificant results are weighted by the new study probabilities in Eq. (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 A.2. If we assume a significance threshold of 5% we obtain:

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

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 Eq. (3.2). For studies in the series later than the pilot study, the expression follows analogously by taking for all $t \geq 2$: $\omega_S^{(t)} = \omega_S$ and $\omega_{NS}^{(t)} = \omega_{NS}$: $\mathbf{E}_0[Z_t \mid T \geq t+1] \approx 1.328$.

To determine the effect on the meta-analysis $Z^{(t)}$-score, we define the expectation under the null hypothesis $\mathbf{E}_0\left[Z^{(t)} \mid T \geq t\right]$, conditioned on the availability of a series of size t. To specify this expectation, we use that the last study is always unbiased since we do not know whether it will spur more studies. As shown in more detail in Appendix A.3, the expression follows from Eq. (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 Eq. (3.4a) and the expression in Eq. (3.4b) under the assumption of equal study sizes ($n_1 = n_2 = \cdots = n_t = n$):

$$\text{For } \alpha = 0.05, \text{ for all } t \geq 2: \quad \mathbf{E}_0\left[Z^{(t)} \mid T \geq t\right] \approx \frac{\sqrt{n_1} \cdot 0.487 + \sum_{i=2}^{t-1} \sqrt{n_i} \cdot 1.328 + \sqrt{n_t} \cdot 0}{\sqrt{N^{(t)}}} \tag{3.4a}$$

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

Table 1 shows the Accumulation Bias in the estimates of $\mathbf{E}_0\left[Z^{(t)} \mid T \geq t\right]$ as studies accumulate under the *Gold Rush* scenario, with equal study sizes and values for the new study probabilities given by Eq. (3.2).

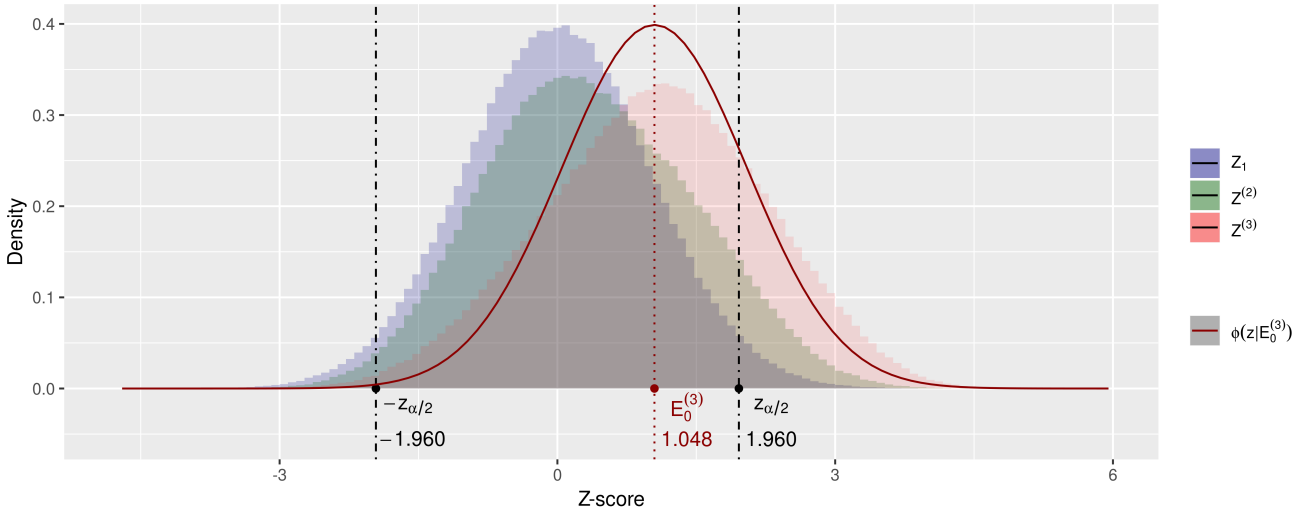

Figure 1. Sampling distributions of meta-analysis $Z^{(t)}$-scores under the null hypothesis in the *Gold Rush* scenario, under the equal study size assumption, with $\alpha = 0.05$ and values for $\omega_s^{(1)}$, $\omega_{NS}^{(1)}$, ω_s and ω_{NS} from Eq. (3.2). $Z^{(t)}$ is as defined in Eq. (3.1b). $\phi(z|\mathrm{E}_0{}^{(3)})$ the standard normal density function shifted by $\mathrm{E}_0{}^{(3)}$, with $\mathrm{E}_0{}^{(3)}$ shorthand for $\mathbf{E}_0\left[Z^{(3)} \middle| T \geq 3\right]$. See Appendix A.7 for the code that produces the simulation and creation of this figure.

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

Figure 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 Eq. (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 1 for a three-study series. Skewness is not the only characteristic that distinguishes the resulting distribution from a standard normal. The variance also 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 A.4. In Figure 1 we see that, with the parameter values from Eq. (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.

3.6 *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 the results are shown in the right hand column of Table 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. Eq. (3.5) expresses the expected type-I error rate for this bias only scenario, with $\Phi()$ the cumulative normal distribution. The inflation actual inflation in the type-I error rate is larger than shown by this scenario, as illustrated the Table 2. The difference between these two type-I error rates for a series of three studies is depicted in Figure 1 by the area under the red histogram for $Z^{(3)}$ and the red $\phi(z \mid \mathrm{E}_0{}^{(3)})$ 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.

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

Table 2. Inflated type-I error rates for tests affected by bias only and tests affected by bias as well as impaired sampling distribution. 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 Eq. (3.2). See Appendix A.7 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

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

We indicated in Section 3.3 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 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.

4 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.

4.1 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 Eq. (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* $A(t)$:

$$A(t \mid z_1, \ldots, z_t) := \mathbf{P}\left[\mathscr{A}^{(t)} \,\middle|\, T \geq t, z_1, \ldots, z_t\right] \cdot S(t-1 \mid z_1, \ldots, z_{t-1}), \tag{4.1}$$

where we define

$$S(t-1 \mid z_1, \ldots, z_{t-1}) := \mathbf{P}[T \geq t \mid z_1, \ldots, z_{t-1}].$$

Eq. (4.1) 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 4.3 and 4.4 we extend the definition and allow analysis time probabilities to also depend on the data in the original scale and external parameters.

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

Just as for the *Gold Rush*' new study probabilities discussed in Section 3.2 and 3.3, 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 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:

[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 $\mathbf{P}[T \geq t \mid z_1, \ldots, z_{t-1}] = \mathbf{P}[T \geq t \mid z_1, \ldots, z_t]$, 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 Eq. (4.1), that $A(t \mid z_1, \ldots, z_t) = \mathbf{P}\left[\mathscr{A}^{(t)}, T \geq t \,\middle|\, z_1, \ldots, z_t\right]$ — so now $A(t \mid z_1, \ldots, z_t)$ is indeed a probability.

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:

$$\lambda(t \mid z_1, \dots, z_t) := \mathbf{P}[T = t \mid T \geq t, z_1, \dots, z_t]. \quad (4.2)$$

In this paper 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 A.2). 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 $t \geq 1$:

$$S(t-1 \mid z_1, \dots, z_{t-1}) = \prod_{i=0}^{t-1} (1 - \lambda(i \mid z_1, \dots, z_i)). \quad (4.3)$$

In many examples, the hazard of stopping at time t, $\lambda(t)$, will depend on the result z_t just obtained. In that case $\lambda(i \mid z_1, \dots, z_i) = \lambda(i \mid z_i)$ in Eq. (4.3) 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.

4.3 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.

4.3.1 *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 Eq. (3.2). The value of $S(t)$, and therefore $A(t)$, can be expressed in terms of these new study probabilities by considering whether $z_1, \dots, z_{t-1}$ 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:

$$\begin{aligned} \text{For } t \text{ such that } \mathbf{P}[\mathscr{A}^{(t)}] = 1: \\ A(t \mid z_1, \dots, z_{t-1}; \alpha) &= S(t-1 \mid z_1, \dots, z_{t-1}; \alpha) \\ &= \prod_{i=0}^{t-1} (1 - \lambda(i \mid z_i; \alpha)), \end{aligned} \quad (4.4)$$

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

$$\begin{aligned} \lambda(i \mid z_i, \alpha) &= 1 - \left(\omega_S^{(i)} \cdot \mathbb{1}_{z_i \geq z_{\frac{\alpha}{2}}} + \omega_{NS}^{(i)} \cdot \mathbb{1}_{|z_i| < z_{\frac{\alpha}{2}}} \right) \\ \overline{\lambda}_0(i \mid \alpha) &:= \mathbf{E}_0[\lambda(i \mid Z_i; \alpha)] \\ &= 1 - \left(\omega_S^{(i)} \cdot \frac{\alpha}{2} + \omega_{NS}^{(i)} \cdot (1 - \alpha) \right). \end{aligned} \quad (4.5)$$

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:

$$\begin{aligned} A(t \mid z_1, \dots, z_{t-1}; \alpha) &= \prod_{i=1}^{t-1} \left(\omega_S^{(i)} \cdot \mathbb{1}_{z_i \geq z_{\frac{\alpha}{2}}} + \omega_{NS}^{(i)} \cdot \mathbb{1}_{|z_i| < z_{\frac{\alpha}{2}}} \right) \\ \overline{A}_0(t \mid \alpha) &:= \mathbf{E}_0[A(t \mid Z_1, \dots, Z_{t-1}; \alpha)] \\ &= \prod_{i=1}^{t-1} \left(\omega_S^{(i)} \cdot \frac{\alpha}{2} + \omega_{NS}^{(i)} \cdot (1 - \alpha) \right). \end{aligned} \quad (4.6)$$

4.3.2 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 size of the series, 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 Section 4.3.4. 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 A.1), 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*:

$$\begin{aligned} &\text{For } t \text{ such that } \mathbf{P}[\mathscr{A}^{(t)}] = 1: \\ &A\left(t \,\middle|\, M^{(t-1)}; \Delta^{H1}, \tau\right) = S\left(t-1 \,\middle|\, M^{(t-1)}; \Delta^{H1}, \tau\right) \\ &\qquad = \prod_{i=0}^{t-1} \left(1 - \lambda\left(i \,\middle|\, M^{(t-1)}; \Delta^{H1}, \tau\right)\right), \end{aligned} \tag{4.7}$$

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

$$\lambda\left(i \,\middle|\, M^{(i-1)}; \Delta^{H1}, \tau\right) = 1 - \left(\frac{M^{(i-1)}}{\Delta^{H1}}\right)^{\tau}, \tag{4.8}$$

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.

4.3.3 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 4.3.1), or on the estimates in the initial studies, such as in the power law model from (Kulinskaya et al., 2016) (Section 4.3.2). (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 7 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.

4.3.4 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 $A(t)$:

$$\begin{aligned} A\left(t \,\middle|\, z^{(1)},\ldots,z^{(t)};z_{\frac{\alpha}{2}}\right) &= \mathbf{P}\left[\mathscr{A}^{(t)} \,\middle|\, T \geq t\right] \\ &\quad \cdot S\left(t-1 \,\middle|\, z^{(1)},\ldots,z^{(t)};z_{\frac{\alpha}{2}}\right) \\ &= S\left(t-1 \,\middle|\, z^{(1)},\ldots,z^{(t-1)};z_{\frac{\alpha}{2}}\right) \\ &= \prod_{i=0}^{t-1}(1-\lambda\left(i \,\middle|\, z^{(i)};z_{\frac{\alpha}{2}}\right)). \end{aligned} \tag{4.9}$$

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:

$$\lambda\left(i \,\middle|\, z^{(i)};z_{\frac{\alpha}{2}}\right) = \mathbb{1}_{|Z^{(i)}|\geq z_{\frac{\alpha}{2}}}. \tag{4.10}$$

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; Thorlund 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, Eq. (4.9) and (4.10) 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 5.2.

4.4 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.

4.4.1 (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 A.1). 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 $A(t)$:

For t such that $S(t-1)=1$:

$$\begin{aligned} A\left(t \,\middle|\, \overline{D}_1,\ldots,\overline{D}_t;a,b\right) &= \mathbf{P}\left[\mathscr{A}^{(t)} \,\middle|\, T \geq t,\overline{D}_1,\ldots,\overline{D}_t;a,b\right] \\ &\quad \cdot S\left(t-1 \,\middle|\, \overline{D}_1,\ldots,\overline{D}_{t-1};a,b\right) \\ &= \mathbf{P}\left[\mathscr{A}^{(t)} \,\middle|\, T \geq t,\overline{D}_1,\ldots,\overline{D}_t;a,b\right] \\ &= \mathbb{1}_{C\in[a,b]} \\ &\quad \text{with} \quad C = \sum_{i=1}^{t} \mathbb{1}_{\overline{D}_i>\Delta^{H1}}. \end{aligned} \tag{4.11}$$

4.5 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 cummulative meta-analyses that show how

Table 3. 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).

	Topics													
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	...	**9 998**	**9 999**	**10 000**
Study series size (t)														
1	$z_{1,1}$	$z_{1,2}$	$z_{1,3}$	$z_{1,4}$	$z_{1,5}$	$z_{1,6}$	$z_{1,7}$	$z_{1,8}$	$z_{1,9}$	$z_{1,10}$	...	$z_{1,9998}$	$z_{1,9999}$	$z_{1,10000}$
2	$z_{2,1}$	$z_{2,2}$	$z_{2,3}$	$z_{2,4}$	$z_{2,5}$		$z_{2,7}$	$z_{2,8}$		$z_{2,10}$		$z_{2,9998}$		$z_{2,10000}$
3	$z_{3,1}$	$z_{3,2}$	$z_{3,3}$		$z_{3,5}$		$z_{3,7}$			$z_{3,10}$		$z_{3,9998}$		$z_{3,10000}$
4		$z_{4,2}$	$z_{4,3}$		$z_{4,5}$		$z_{4,7}$					$z_{4,9998}$		$z_{4,10000}$
5		$z_{5,2}$			$z_{5,5}$							$z_{5,9998}$		
6		$z_{6,2}$			$z_{6,5}$							$z_{6,9998}$		
...												...		
136												$z_{136,9998}$		

many studies were still performed while satisfactory evidence was already available. These cummulative meta-analysis 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 (Section 4.3.4).

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 (Section 4.3.2) — 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, we do not need to be that pessimistic and can still embrace *Evidence-Based Research*. In Section 7 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.

5 *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 Table 3.

The conventional approach to error control, which we used to show the influence of *Gold Rush* Accumulation Bias in meta-analysis testing in Section 3.6, 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 4.3.1 we took the t at which the sole meta-analysis is conducted to be either random or prespecified. This is shown in Table 3 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 Table 3 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 *Toy Story* 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 4.4. We discuss the conventional conditional approach to meta-analysis error control in more detail in Section 5.1.

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 Table 3 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 5.2.

5.1 Error control conditioned on time

The null distributions of the common/fixed meta-analysis Z-statistic shown in Figure 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 $f_0(z^{(t)})$ to denote the assumed standard normal null distribution for the meta-analysis Z-score and obtain a conditional density using Bayes' rule:

$$f_0\left(z^{(t)} \,\middle|\, \mathscr{A}^{(t)}, T \geq t\right) = \frac{f_0(z^{(t)}) \cdot \mathbf{P}_0\left[\mathscr{A}^{(t)}, T \geq t \,\middle|\, z^{(t)}\right]}{\mathbf{P}_0\left[\mathscr{A}^{(t)}, T \geq t\right]} = \frac{f_0(z^{(t)}) \cdot \bar{A}_0\left(t \,\middle|\, z^{(t)}\right)}{\bar{A}_0(t)},$$

where we define:

$$\bar{A}_0\left(t \,\middle|\, z^{(t)}\right) := \mathbf{E}_0\left[A(t \,|\, Z_1, \ldots, Z_t) \,\middle|\, Z^{(t)} = z^{(t)}\right]$$
$$\bar{A}_0(t) := \mathbf{E}_0\left[A(t \,|\, Z_1, \ldots, Z_t)\right],$$

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

$$Z^{(t)} = \frac{1}{\sqrt{t}} \sum_{i=1}^{t} Z_i \tag{5.1}$$

(extension to the general cases with unequal sample sizes is straightforward). For the *Gold Rush* example, $\bar{A}_0(t)$ was given by Eq. (4.6) and can be calculated if ωs are known. $\bar{A}_0(t)$ 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 2 are based on a randomly chosen or prespecified t for which $\mathbf{P}[\mathscr{A}^{(t)}] = 1$, and represent the following (with f_0 as above in Eq. (5.1)):

$$\mathbf{P}_0\left[\mathscr{E}^{(t)}_{\text{TYPE-I}} \,\middle|\, \mathscr{A}^{(t)}, T \geq t\right] = \int_{-\infty}^{-z_{\frac{\alpha}{2}}} f_0\left(z^{(t)} \,\middle|\, \mathscr{A}^{(t)}, T \geq t\right) dz^{(t)} + \int_{z_{\frac{\alpha}{2}}}^{\infty} f_0\left(z^{(t)} \,\middle|\, \mathscr{A}^{(t)}, T \geq t\right) dz^{(t)}. \tag{5.2}$$

5.2 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:

$$f_0\left(z^{(1)}, \ldots, z^{(t)}, T = t\right) = f_0\left(z^{(1)}, \ldots, z^{(t)}\right) \cdot \mathbf{P}_0\left[T = t \,\middle|\, z^{(1)}, \ldots, z^{(t)}\right], \tag{5.3}$$

where we define

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

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

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

and with $f_0(z^{(0)}) = 1$ 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 A.5. 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; Thorlund 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:

$$\mathbf{E}_T\left[\mathbf{P}_0\left[\mathscr{E}^{(T)}_{\text{TYPE-I}} \,\middle|\, T\right]\right] = \sum_{t=1}^{T_{\max}} \int_{z^{(1)}_{\frac{\alpha}{2}}}^{\infty} \cdots \int_{z^{(t)}_{\frac{\alpha}{2}}}^{\infty} f_0\left(z^{(1)}, \ldots, z^{(t)}, T = t\right) dz^{(1)} \ldots dz^{(t)} = \alpha,$$

with f_0 as above (5.3)
and $T = t$ only in the case $\lambda(t) = \mathbb{1}_{Z^{(t)} \geq z^{(t)}_{\frac{\alpha}{2}}} = 1$. (5.4)

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.

5.3 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 Sections 4.3 and 4.4 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 4.4, 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 Table 3 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 7, 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.

6 Intermezzo: evidence for the existence of Accumulation Bias

6.1 Agreement with empirical findings

Accumulation Bias arises due to dependencies in how a study series comes about (Section 4.3), and in the timing of the meta-analysis (Section 4.4). 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) (Section 4.3.3). 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 extend 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 this 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) (Section 4.3.3), might therefore be a too 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.

6.2 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 fraction of pilot studies from the null and alternative distribution ($\pi/(1-\pi)$) in the terminology of this paper. 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. This is the expected rate or target rate of true to false rejections, and the same as $\gamma \cdot \pi/(1-\pi)$ in Section 7.1 of this paper.

Ioannidis (2005b) provides prior odds of various research fields and publication types for which two 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 $\frac{\pi \cdot \bar{A}_1(t)}{(1-\pi)\cdot \bar{A}_0(t)}$, with $\bar{A}_1(1) = 1$ and $\bar{A}_0(1) = 1$ for primary studies.

7 Likelihood ratios' independence from meta-analysis time

In Section 5.3 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 7.1 from the perspective of error control conditioned on time, and in Section 7.2 from the perspective of error control surviving over time.

Our proposed use of the likelihood ratio is based on the following extraordinary property[2], already recognized by Berger and Berry (1988) and shown in Eq. (7.1): The likelihood ratio is a test statistic that depends on the specification of some alternative distribution f_1. 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 4.2). 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:

$$\begin{aligned}
&\mathbf{LR}_{10}{}^{(t)}\left(z_1,\ldots,z_t,\mathscr{A}^{(t)},T\geq t\right)\\
&:=\frac{f_1(z_1,\ldots,z_t)\cdot\mathbf{P}_1(\mathscr{A}^{(t)},T\geq t\mid z_1,\ldots,z_t)}{f_0(z_1,\ldots,z_t)\cdot\mathbf{P}_0(\mathscr{A}^{(t)},T\geq t\mid z_1,\ldots,z_t)}\\
&=\frac{f_1(z_1,\ldots,z_t)\cdot A(t\mid z_1,\ldots,z_t)}{f_0(z_1,\ldots,z_t)\cdot A(t\mid z_1,\ldots,z_t)}\\
&=\frac{f_1(z_1,\ldots,z_t)}{f_0(z_1,\ldots,z_t)}\\
&=\mathbf{LR}_{10}(z_1,\ldots,z_t).
\end{aligned}\tag{7.1}$$

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]

7.1 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 Table 3) π are sampled from the alternative distribution and a proportion $(1-\pi)$ 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 $\mathbf{P}_0\left[\mathscr{E}^{(t)}_{\text{TYPE-I}}\,\middle|\,\mathscr{A}^{(t)}, T\geq t\right]$.

[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 gave rise to the observations (Hendriksen et al., 2018)

[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 $\mathbf{LR}_{10}(z_1,\ldots,z_t)\geq\gamma$ where γ is chosen such that $\mathbf{P}_0(f_1(z_1,\ldots,z_t)/f_0(z_1,\ldots,z_t)\geq\gamma)=\alpha$. 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 7.2 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 $O_{\text{post}}(z_1,\ldots,z_t) = \mathbf{LR}_{10}(z_1,\ldots,z_t)\cdot\frac{\pi}{1-\pi}$. Remarkably, these are not affected by the mechanism underlying the decisions to continue studies or perform meta-analyses:

$$\begin{aligned} & O_{\text{post}}\left(z_1,\ldots,z_t \,\middle|\, \mathscr{A}^{(t)}, T \geq t\right) \\ & := \frac{\mathbf{P}\left[H_1 \,\middle|\, z_1,\ldots,z_t, \mathscr{A}^{(t)}, T \geq t\right]}{\mathbf{P}\left[H_0 \,|\, z_1,\ldots,z_t, \mathscr{A}^{(t)}, T \geq t\right]} \\ & = \frac{f_1\left(z_1,\ldots,z_t, \mathscr{A}^{(t)}, T \geq t\right)\cdot\pi}{f_0\left(z_1,\ldots,z_t, \mathscr{A}^{(t)}, T \geq t\right)\cdot(1-\pi)} \\ & = \mathbf{LR}_{10}{}^{(t)}\left(z_1,\ldots,z_t, \mathscr{A}^{(t)}, T \geq t\right)\cdot\frac{\pi}{1-\pi} \\ & = \mathbf{LR}_{10}(z_1,\ldots,z_t)\cdot\frac{\pi}{1-\pi} \\ & = O_{\text{post}}(z_1,\ldots,z_t). \end{aligned} \tag{7.2}$$

We can set a threshold γ based on the rate of true to false rejections, so $\gamma = 16$ would mean that we try to achieve 16 times as many true rejections than false rejections $\gamma = \frac{1-\beta}{\alpha}$, which is the 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 the *pre-experimental rejection odds* (Bayarri et al., 2016) $\gamma\cdot\frac{\pi}{1-\pi}$ and use these to threshold the posterior odds (Eq. (7.2)). We define R to be the region of the sample space and $\mathscr{R}$ the event for which $O_{\text{post}}(z_1,\ldots,z_t) \geq \gamma\cdot\frac{\pi}{1-\pi}$, i.e. the event that we reject, and obtain the following:

$$\begin{aligned} & \frac{1-\mathbf{P}_1\left[\mathscr{E}^{(t)}_{\text{TYPE-II}} \,\middle|\, \mathscr{A}^{(t)}, T \geq t\right]}{\mathbf{P}_0\left[\mathscr{E}^{(t)}_{\text{TYPE-I}} \,\middle|\, \mathscr{A}^{(t)}, T \geq t\right]} \\ & = \frac{\mathbf{P}_1\left[O_{\text{post}}\left(Z_1,\ldots,Z_t \,\middle|\, \mathscr{A}^{(t)}, T \geq t\right) \geq \gamma\cdot\frac{\pi}{1-\pi}\right]}{\mathbf{P}_0\left[O_{\text{post}}\left(Z_1,\ldots,Z_t \,|\, \mathscr{A}^{(t)}, T \geq t\right) \geq \gamma\cdot\frac{\pi}{1-\pi}\right]} \\ & = \frac{\mathbf{P}_1\left[O_{\text{post}}(Z_1,\ldots,Z_t) \geq \gamma\cdot\frac{\pi}{1-\pi}\right]}{\mathbf{P}_0\left[O_{\text{post}}(Z_1,\ldots,Z_t) \geq \gamma\cdot\frac{\pi}{1-\pi}\right]} \\ & = \frac{\mathbf{P}_1[\mathscr{R}]}{\mathbf{P}_0[\mathscr{R}]} \geq \frac{\mathbf{P}_1[\mathscr{R}]}{\mathbf{P}_1[\mathscr{R}]\cdot\frac{1}{\gamma}} = \gamma, \end{aligned} \tag{7.3}$$

where the inequality follows since if $O_{\text{post}}\left(z_1,\ldots,z_t \,\middle|\, \mathscr{A}^{(t)}, T \geq t\right) \geq \gamma\cdot\frac{\pi}{1-\pi}$:

$$\begin{aligned} & \frac{f_1(z_1,\ldots,z_t)}{f_0(z_1,\ldots,z_t)}\cdot\frac{\pi}{1-\pi} \geq \gamma\cdot\frac{\pi}{1-\pi} \\ & \text{then} \quad \frac{f_1(z_1,\ldots,z_t)}{f_0(z_1,\ldots,z_t)} \geq \gamma \quad \text{and} \\ & \mathbf{P}_0[\mathscr{R}] = \int_R f_0(z_1,\ldots,z_2) \leq \int_R \frac{f_1(z_1,\ldots,z_2)}{\gamma} = \frac{\mathbf{P}_1[\mathscr{R}]}{\gamma}. \end{aligned} \tag{7.4}$$

So by specifying $\frac{\pi}{1-\pi}$ and an intended rate of true to false rejections γ, we can calculate the posterior odds based on the likelihood ratio, compare it to the threshold based on γ and control fraction γ 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.

7.2 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 5.2. 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:

$$\begin{aligned} & \mathbf{P}_0\left[\text{there exists } t \leq T \text{ with } \mathscr{E}^{(t)}_{\text{TYPE-I}} \text{ and } \mathscr{A}^{(t)}\right] \\ & = \mathbf{P}_0\left[\exists t \leq T:\ \mathscr{E}^{(t)}_{\text{TYPE-I}};\mathscr{A}^{(t)}\right] \\ & = \mathbf{P}_0\left[\exists t \leq T:\ \mathbf{LR}_{10}{}^{(t)}(Z_1,\ldots,Z_t) \geq \frac{1}{\alpha};\mathscr{A}^{(t)}\right] \\ & \leq \mathbf{P}_0\left[\exists t > 0:\ \mathbf{LR}_{10}{}^{(t)}(Z_1,\ldots,Z_t) \geq \frac{1}{\alpha}\right] \leq \alpha. \end{aligned} \tag{7.5}$$

The final inequality is a classic result, proofs of which can be found in, for example, Robbins (1970); Shafer et al. (2011) and (with substantial explanation) Hendriksen et al. (2018); 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 A.6, the previous inequality implies that

$$\sum_t \mathbf{P}_0\left[\mathscr{F}^{(t)}_{\text{TYPE-I}}, \mathscr{A}^{(t)}, T \geq t\right] \leq \alpha. \tag{7.6}$$

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 Section 4.3.4). 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 Eq. (7.6). 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 Eq. (7.6) 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 $A(t \mid Z_1, \ldots, Z_t)$ ($\bar{A}_0(t)$) 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 Table 3, 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.

8 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 π of true alternative hypotheses among pilot studies to threshold the posterior odds. Assuming a base rate π 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 π, to see how it affects our posterior odds. The important thing for us to note within the context of this paper 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.

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.

9 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.

[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)

Fortunately, much simpler approaches to the same goal can be found. One intuitive way is to consider a series of bets $s(Z_1), s(Z_2), \ldots, s(Z_t)$ against the null hypothesis 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 might costs \$1 to play the bet, but in expectation also makes a \$1 profit:

$$\mathbf{E}_0[s(Z_t)] = \$1. \tag{9.1}$$

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 Eq. (9.1) 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 $s(Z_t) = f_1(Z_t)/f_0(Z_t)$, then Eq. (9.1) obviously holds:

$$\mathbf{E}_0\left[\frac{f_1(Z_t)}{f_0(Z_t)}\right] = \int_z f_0(z)\frac{f_1(z)}{f_0(z)}dz = \int_z f_1(z)dz = 1. \tag{9.2}$$

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 Eq. (7.5) and (7.6) — 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 under the null 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 is a way to test the data without loosing the ability to build on it in future studies. The likelihood ratio has the ability to maximize the rate of growth 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, 2019). 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).

9.1 Expanding likelihood ratios to *Safe Tests*

When the null hypothesis is simple, it can be shown that either using bets that satisfy Eq. (9.1) 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 Eq. (9.1) 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 7.2) can still be obtained for general composite null.

10 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 that we will study in a forthcoming paper. 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, 2019): 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 π and separate meta-analysis 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 paper 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)

Data availability

Underlying data

All data underlying the results are available as part of the article and no additional source data are required

Extended data

See Appendix A.7 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

(Schure, 2019)

Grant information

This work is part of the NWO TOP-I research programme *Safe Bayesian Inference* [617.001.651], which is financed by the Netherlands Organisation for Scientific Research (NWO).

Acknowledgements

This paper benefited from discussions with Allard Hendriksen, Rosanne Turner, Muriel Pérez, Alexander Ly and Glenn Shafer.

References

Armitage, P. (1984). Controversies and achievements in clinical trials. *Contemporary Clinical Trials*, 5(1):67–72.

Bayarri, M., Benjamin, D. J., Berger, J. O., and Sellke, T. M. (2016). Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology*, 72:90–103.

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.

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., 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.

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.

Egger, M. and Smith, G. D. (1998). Bias in location and selection of studies. *BMJ: British Medical Journal*, 316(7124):61.

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.

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.

Gøtzsche, P. C. (1987). Reference bias in reports of drug trials. *Br Med J (Clin Res Ed)*, 295(6599):654–656.

Grünwald, P. D., De Heide, R., and Koolen, W. (2019). Safe testing. *arXiv preprint*.

Hendriksen, A., de Heide, R., and Grünwald, P. (2018). Optional stopping with Bayes factors: a categorization and extension of folklore results, with an application to invariant situations. *arXiv preprint arXiv:1807.09077*.

Higgins, J., Whitehead, A., and Simmonds, M. (2011). Sequential methods for random-effects meta-analysis. *Statistics in medicine*, 30(9):903–921.

Ioannidis, J. (2010). Meta-research: The art of getting it wrong. *Research Synthesis Methods*, 1(3-4):169–184.

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. 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.

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.

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.

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.

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.

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.

Proschan, M. A., Lan, K. G., and Wittes, J. T. (2006). *Statistical monitoring of clinical trials: a unified approach*. Springer Science & Business Media.

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.

Royall, R. (2000). On the probability of observing misleading statistical evidence. *Journal of the American Statistical Association*, 95(451):760–768.

Schure, J. T. (2019). Accumulation bias in meta-analysis: The need to consider time in error control.

Shafer, G. (2019). The language of betting as a strategy for statistical and scientific communication. http://probabilityandfinance.com/articles/54.pdf. Online; accessed 16 May 2019.

Shafer, G., Shen, A., Vereshchagin, N., Vovk, V., et al. (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*. Wiley.

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.

Thorlund, K., Devereaux, P., Wetterslev, J., Guyatt, G., Ioannidis, J. P., Thabane, L., Gluud, L.-L., Als-Nielsen, B., and Gluud, C. (2008). Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses? *International journal of epidemiology*, 38(1):276–286.

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.

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.

A Appendix

A.1 Common/fixed-effect meta-analysis

Here we derive Eq. (3.1a) and (3.1b), shown in (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:

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

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. Equation (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)):

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

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 $\overline{D}_i = d_i \sigma_D$, $SE^2_{\overline{D}_i} = \frac{\sigma^2_D}{n_i}$, and thus $W_i = w_i \frac{1}{\sigma^2_D}$ (see also Borenstein et al. (2009, p. 82)):

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

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:

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

A.2 Expectation *Gold Rush* conditional pilot Z-score

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

$$\begin{aligned}\mathbf{E}_0[Z_1 \mid T \geq 2] &= \frac{\mathbf{E}_0\left[Z_1 \,\middle|\, T \geq 2, Z_1 \geq z_{\frac{\alpha}{2}}\right] \cdot \mathbf{P}_0\left[T \geq 2 \,\middle|\, T \geq 1, Z_1 \geq z_{\frac{\alpha}{2}}\right] \cdot \mathbf{P}_0\left[Z_1 \geq z_{\frac{\alpha}{2}}\right]}{\mathbf{P}_0[T \geq 2]} \\ &\quad + \frac{\mathbf{E}_0\left[Z_1 \,\middle|\, T \geq 2, |Z_1| < z_{\frac{\alpha}{2}}\right] \cdot \mathbf{P}_0\left[T \geq 2 \,\middle|\, T \geq 1, |Z_1| < z_{\frac{\alpha}{2}}\right] \cdot \mathbf{P}_0\left[|Z_1| < z_{\frac{\alpha}{2}}\right]}{\mathbf{P}_0[T \geq 2]} \\ &= \frac{\mathbf{E}_0\left[Z_1 \,\middle|\, T \geq 2, Z_1 \geq z_{\frac{\alpha}{2}}\right] \cdot \omega_{\mathrm{S}}^{(1)} \cdot \frac{\alpha}{2} + \mathbf{E}_0\left[Z_1 \,\middle|\, T \geq 2, |Z_1| < z_{\frac{\alpha}{2}}\right] \cdot \omega_{\mathrm{NS}}^{(1)} \cdot (1-\alpha)}{\omega_{\mathrm{S}}^{(1)} \cdot \frac{\alpha}{2} + \omega_{\mathrm{NS}}^{(1)} \cdot (1-\alpha)}\end{aligned} \tag{A.5}$$

since

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

This expression only considers significant positive and nonsignificant results in the pilot study, since we defined in Eq. (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 3.3.

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

For all $t \geq 2$:

$$\begin{aligned}\mathbf{E}_0\left[Z^{(t)} \,\middle|\, T \geq t\right] &= \frac{\sum_{i=1}^{t} \sqrt{n_i}\, \mathbf{E}_0[Z_i \mid T \geq t]}{\sqrt{N^{(t)}}} \\ &= \frac{\sqrt{n_1}\, \mathbf{E}_0[Z_1 \mid T \geq t] + \sum_{i=2}^{t-1} \sqrt{n_i}\, \mathbf{E}_0[Z_i \mid T \geq t] + \sqrt{n_t}\, \mathbf{E}_0[Z_t \mid T \geq t]}{\sqrt{N^{(t)}}} \\ &= \frac{\sqrt{n_1}\, \mathbf{E}_0[Z_1 \mid T \geq 2] + \sum_{i=2}^{t-1} \sqrt{n_i}\, \mathbf{E}_0[Z_i \mid T \geq i+1]}{\sqrt{N^{(t)}}}\end{aligned} \tag{A.6}$$

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 $\mathbf{E}_0[Z_i \mid i+1]$).

A.4 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] \\
&\quad\quad - \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) \\
&\quad\quad + (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) \\
&\quad\quad - \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\} \\
&\quad\quad + \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{(A.7a)} \\
&\quad\quad - \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{(A.7b)}
\end{aligned}
$$

Because squaring is a convex function, we know from Jensen's Inequality that the average squared mean (A.7a) is larger than the square of the average mean (A.7b). So the variance of the mixture is larger than the mixture of the variances.

A.5 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* $F(t-1)$:

$$
\begin{aligned}
S(t-1) &= 1 - F(t-1) \\
\text{with} \quad F(t-1) &= \sum_{i=0}^{t-1} \mathbf{P}[T = i]
\end{aligned}
\qquad \text{(A.8)}
$$

$$
\begin{aligned}
S(t-1) &= 1 - \sum_{i=0}^{t-1} \mathbf{P}[T = i] \\
S(t) &= 1 - \sum_{i=0}^{t-1} \mathbf{P}[T = i] - \mathbf{P}[T = t] \\
\text{therefore:} \quad \mathbf{P}[T = t] &= S(t-1) - S(t)
\end{aligned}
\qquad \text{(A.9)}
$$

A.6 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:

$$
\begin{aligned}
\sum_t \mathbf{P}_0\left[\mathscr{F}^{(t)}_{\text{TYPE-I}}, \mathscr{A}^{(t)}, T \geq t\right] &\leq \sum_t \mathbf{P}_0\left[\mathscr{F}^{(t)}_{\text{TYPE-I}}, T \geq t\right] \\
&= \mathbf{P}_0\left[\exists t > 0 : \mathscr{F}^{(t)}_{\text{TYPE-I}}, T \geq t\right] \\
&\leq \mathbf{P}_0\left[\exists t > 0 : \mathscr{F}^{(t)}_{\text{TYPE-I}}\right] \\
&= \mathbf{P}_0\left[\exists t > 0 : \mathscr{E}^{(t)}_{\text{TYPE-I}}\right] \\
&= \mathbf{P}_0\left[\exists t > 0 : \mathbf{LR}_{10}{}^{(t)}(Z_1, \ldots, Z_t) \geq \frac{1}{\alpha}\right] \leq \alpha
\end{aligned}
$$

where the final inequality is just the final inequality of (7.5) again. (7.6) follows.

A.7 Code availability

Table 1, Figure 1 and Table 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(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