Accumulation Bias: How to handle it ALL-IN

Judith ter Schure

February 10, 2021

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

Ter Schure and Grünwald (2019) detail all the possible ways in which the size of a study series up for meta-analysis, or the timing of the meta-analysis, might be driven by the results within those studies. Any such dependency introduces *accumulation bias*. Unfortunately, it is often impossible to fully characterize the processes at play in retrospective meta-analysis. The bias cannot be accounted for. In this blog we revisit an example accumulation bias process, that can be one of many influencing a single meta-analysis, and use it to illustrate the following key points:

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

Our example: extreme *Gold Rush* accumulation bias

We imagine a world in which a series of studies is meta-analyzed as soon as three studies become available. Many topics deserve a first initial study, but the research field is very selective with its replications. Nevertheless, for significant results in the right direction, a replication is warranted. We call this the *Gold Rush* scenario, because after each finding of a positive significant result – the gold in science – some research group rushes into a replication, but as soon as a study disappoints, the research effort is terminated and no-one bothers to ever try again. This scenario was first proposed by Ellis and Stewart (2009) and formulated in detail and under this name by Ter Schure and Grünwald (2019). Here we consider the most extreme version of the *Gold Rush* where finding a significant positive result not only makes a replication more probable, but even inevitable: the dependency of occurring replications on their predecessor's result is deterministic.

Biased *Gold Rush* sampling

We denote the number of studies available on a certain topic by t. This number t can also indicate the *timing* of a meta-analysis, such that a meta-analysis can possibly occur at number of studies $t = 1, 2, 3, \ldots$ up to some maximum number of studies T. This notation follows from Ter Schure and Grünwald (2019); the Technical Details at the end of this blog make the notation involved in this blog more explicit.

We summarize the results of individual studies into a single per-study Z-score (z_1 for the first study, z_2 for the second, etc), such that we have the following information on a series of size t:

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

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

Gold Rush world

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

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

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

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

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

Assume that we are in the scenario that only true null effects are studied in our *Gold Rush* world, such that any new study builds on a false-positive result. How large would the bias be if the three-study series are simply analyzed by standard meta-analysis? We illustrate this by simulating this *Gold Rush* world using the R code below.

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

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

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

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

# meta Zscores for a random sample of 3-study series
Zmeta3 <- sapply(sample(1:numSim.study, size = numSim.3series), function(i) calcZmeta(c(Z1[i], Z2[i], Z3[i])))

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

dataZmeta.cond <- data.frame(Zscore = c(Zmeta3, Zmeta3.A3),
                             GoldRush = c(rep("Zmeta3", times = numSim.3series),
                                          rep("Zmeta3.A3", times = numSim.3series)))
ggplot(dataZmeta.cond) +
  geom_histogram(aes(x = Zscore, y = ..density.., fill = GoldRush),
                 alpha = 0.2, bins = 120, position = "identity") +
  scale_fill_discrete(name = "Gold Rush A(t)",
                      labels = c(expression(z^(3)),
                                 expression(paste("A(3) = 1     ", z^(3)))))
```

Figure 1. Code to create Figure 2

Theoretical sampling process: A fixed-effects meta-analysis assumes that if three studies z_1, z_2, z_3 are each sampled under the null hypothesis, each has a standard normal with mean zero and the standard normal sampling distribution also applies for the combined $z^{(3)}$ score. The R code in Figure 1 illustrates this sampling process: First, a large population is simulated of possible first (`Z1`), second (`Z2`) and third (`Z3`) studies from a standard normal distribution. Then in `Zmeta3` each index `i` represents a possible study series, such that `c(Z1[i], Z2[i], Z3[i])` samples an unbiased study series and `calcZmeta` calculates its fixed-effects meta-analysis Z-score $z^{(3)}$. So the large number of Z-scores in `Zmeta3` captures the unbiased sampling distribution that is assumed for fixed-effects meta-analysis $z^{(3)}$-scores.

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

Meta-analysis under *Gold Rush* accumulation bias: The final lines of code in Figure 1 plot two histograms of $z^{(3)}$ samples, one with and one without the *Gold Rush* $A(t)$ accumulation bias process, based on `Zmeta3.A3` and `Zmeta3` respectively. Figure 2 gives the result.

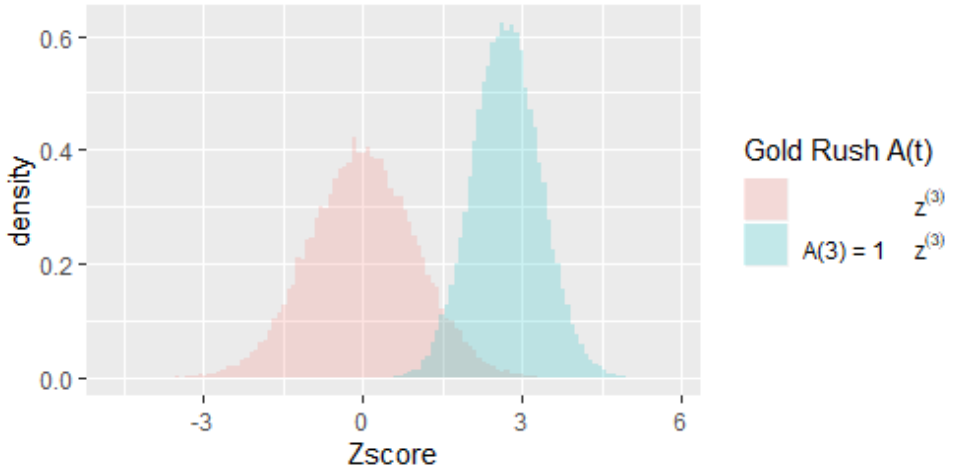

Figure 2. Sampling distributions under the null hypothesis of fixed-effects meta-analysis Z-scores $Z^{(3)}$ of three studies with and without extreme *Gold Rush* accumulation bias $A(t)$, under the assumption of equal study sample size and variance.

We observe in Figure 2 that the theoretical sampling process, resulting in the pink histogram, gives a distribution for the three-study meta-analysis $z^{(3)}$-scores that is centered around zero. Under the *Gold Rush* sampling process, however, our three-study $z^{(3)}$-scores do not behave like this theoretical distribution at all. The blue histogram has a smaller variance and is shifted to the right – representing the bias.

We conclude that we should not use conventional meta-analysis techniques to analyze our study series under *Gold Rush* accumulation bias: Conventional fixed-effects meta-analysis assumes that any three-study summary statistic $Z^{(3)}$ is sampled from the pink distribution in Figure 2 under the null hypothesis, such that the meta-analysis is significant for $Z^{(3)}$-scores larger than $z_\alpha = 1.96$ for a right-sided test with type-I error control $\alpha = 2.5\%$. Yet the actual blue sampling distribution under this accumulation bias process shows that a much larger fraction of series that accumulate three studies will have $Z^{(3)}$-scores larger than 1.96 than is assumed by the theory of random sampling. This (extremely) inflated proportion of type-I errors is 88% instead of 2.5% in our extreme *Gold Rush*, and can be obtained from our simulation by the code in Figure 3.

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

Figure 3. Code to calculate type-I error probability with and without extreme *Gold Rush* accumulation bias.

Accumulation bias can be efficient

The steps in the code from Figure 1 that arrive at the biased distribution in Figure 2 illustrate that accumulation bias is in fact a selection bias. Nevertheless, accumulation bias does not result from questionable research practices, such as publication bias from file-drawering a selection of results. The selection to replicate only some studies instead of all of them biases the sampling distribution of study series, but can be a very efficient approach to set priorities in research and reduce research waste.

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

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

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

Gold Rush world; all-series-size

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

The R code in Figure 4 calculates the fixed-effects meta-analysis $z^{(1)}$, $z^{(2)}$ and $z^{(3)}$ scores, conditional on meta-analyzing a one-study, two-study, or three-study series in this adjusted *Gold Rush* accumulation bias scenario. The histograms of these conditional $z^{(t)}$ scores are shown in Figure 5, including the theoretical unbiased $z^{(3)}$ histogram that was also shown in Figure 2 and largely overlaps with the "$A(1) = 1, A(2) = 0$"-scenario. The difference between these two sampling distributions is only visible in their right tail, with the green histogram excluding values larger than $z_\alpha = 1.96$ and redistributing their mass over other values.

Figure 5 clarifies that single studies are hardly biased in this extreme *Gold Rush* scenario, that the bias is problematic for two-study series and most extreme for three-study ones. However, what this plot does not show us is how often we are in the one-study, two-study and three-study case.

To illustrate the relative frequencies of one-study, two-study and three-study meta-analyses, the code in Figure 6 samples the series in their respective numbers, instead of in equal numbers (which happens in the `size = numSim.3series` statement in Figure 4, part of creating the data frame). Plotting the total number of sampled

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

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

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



dataZmeta.cond <- rbind(dataZmeta.cond,
                        data.frame(Zscore = c(sample(Zmeta1.A1notA2, size = numSim.3series),
                                              sample(Zmeta2.A2notA3, size = numSim.3series)),
                                   GoldRush = c(rep("Zmeta1.A1notA2", times = numSim.3series),
                                                rep("Zmeta2.A2notA3", times = numSim.3series))))
dataZmeta.cond$GoldRush <- factor(dataZmeta.cond$GoldRush,
                                  levels = c("Zmeta3", "Zmeta1.A1notA2", "Zmeta2.A2notA3", "Zmeta3.A3"))
ggplot(dataZmeta.cond) +
  geom_histogram(aes(x = Zscore, y = ..density.., fill = GoldRush),
                 alpha = 0.2, bins = 120, position = "identity") +
  scale_fill_discrete(name = "Gold Rush A(t)",
                      labels = c(expression(z^(3)),
                                 expression(paste("A(1) = 1, A(2) = 0     ", z^(1))),
                                 expression(paste("A(2) = 1, A(3) = 0     ", z^(2))),
                                 expression(paste("A(3) = 1     ", z^(3)))))
```

Figure 4. Code to create Figure 5

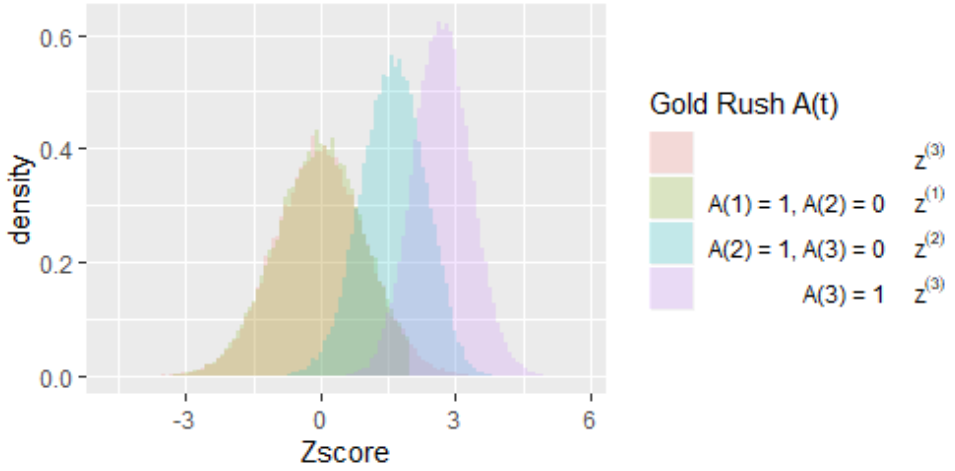

Figure 5. Sampling distributions under the null hypothesis of fixed-effects meta-analysis Z-scores $Z^{(t)}$ of one, two and three studies with extreme *Gold Rush* accumulation bias and a three-study meta-analysis without accumulation bias, under the assumption of equal study sample size and variance.

Z-scores is dangerous for the single study $z^{(1)}$-scores, however, since there are so many of them (it can crash your R studio). So before plotting the histogram, a smaller sample (of `size = 3*numSim.3series` in total) is drawn that keeps the ratios between $z^{(1)}$s, $z^{(2)}$s and $z^{(3)}$s intact.

The histogram in Figure 7 illustrates an unconditional distribution by the raw counts of the $z^{(t)}$-scores: many result from a single study, very few from a two-study series and almost none from a three-study series. In fact, this unconditional sampling distribution is hardly biased, as we will illustrate with our table further below.

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

```r
dataZmeta.unc <- data.frame(Zscore = c(Zmeta1.A1notA2,  # almost 64 million samples, beware in histogram
                                       Zmeta2.A2notA3, Zmeta3.A3),
                            GoldRush = c(rep("Zmeta1.A1notA2", times = length(A1notA2)),
                                         rep("Zmeta2.A2notA3", times = length(A2notA3)),
                                         rep("Zmeta3.A3", times = numSim.3series)))


ggplot(dataZmeta.unc[sample(1:nrow(dataZmeta.unc), size = 3*numSim.3series), ]) + # histogram << 64 million
  geom_histogram(aes(x = Zscore, y = ..count.., fill = GoldRush),
                 alpha = 0.2, bins = 120, position = "identity") +
  scale_fill_discrete(name = "Gold Rush A(t)",
                      labels = c(expression(paste("A(1) = 1, A(2) = 0     ", z^(1))),
                                 expression(paste("A(2) = 1, A(3) = 0     ", z^(2))),
                                 expression(paste("A(3) = 1     ", z^(3)))))
```

Figure 6. Code to create Figure 7

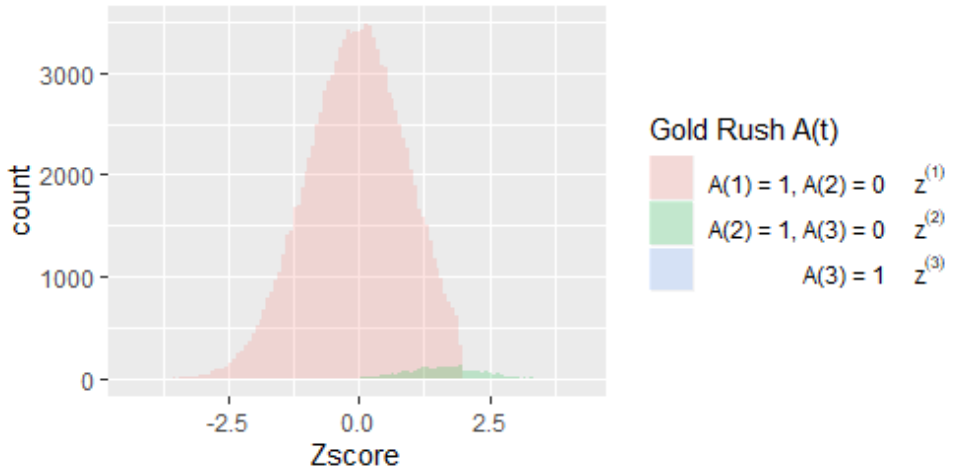

Figure 7. Unconditional sampling distributions under the null hypothesis of fixed-effects meta-analysis Z-scores $Z^{(t)}$ of either one, two or three studies under extreme *Gold Rush* accumulation bias, under the assumption of equal study sample size and variance.

ALL-IN meta-analysis

Figure 8 shows an example of an ALL-IN meta-analysis. Each of the red/orange/yellow lines represents a study out of the ten separate studies in as many different countries. The blue line indicates the meta-analysis synthesis of the evidence; a live account of the evidence so far in the underlying studies. In fact, *ALL-IN* meta-analysis stands for *Anytime, Live* and *Leading INterim* meta-analysis, in which the *Anytime Live* property assures valid inference under continuously monitoring and the *Leading* property allows the meta-analysis results to inform whether individual studies should be stopped or expanded. It should be noted that such data-driven decisions would invalidate conventional meta-analysis by introducing accumulation bias.

To interpret Figure 8, we observe that initially only the Australian (AU) study contributes to the meta-analysis and the blue line completely overlaps with the red one. Very quickly, the Dutch (NL) study also starts contributing and the blue meta-analysis line captures a synthesis of the evidence in two studies. Later on, also the study in the US, France (FR) and Uruguay (UY) start contributing and the meta-analysis becomes a three-study, four-study and five-study meta-analysis. How many studies contribute to the analysis, however, does not matter for its evidential value. Some studies (like the Australian one) are much larger than others, such that under a lucky scenario this study could reach the evidential threshold even before other studies start observing data. This threshold (indicated at 400) controls type-I errors at a rate of $\alpha = 1/400 = 0.0025$ (details in the final section). So in repeated sampling under the null, the combined studies will only have a probability

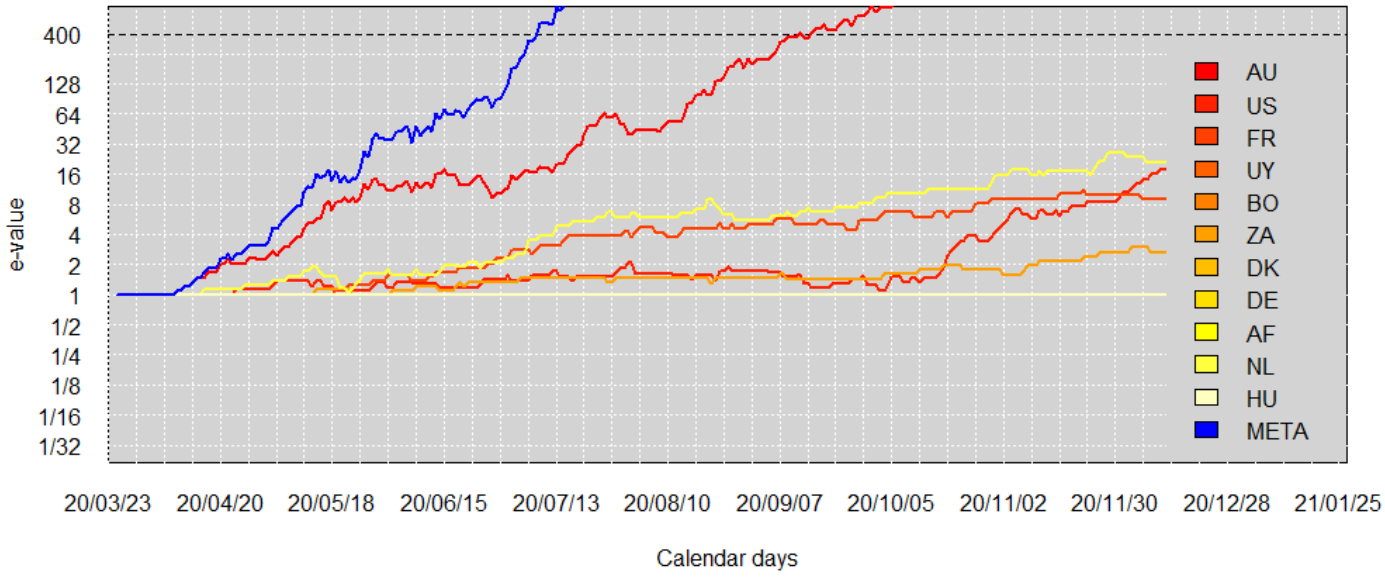

Figure 8. Dashboard of an ALL-IN meta-analysis of between one and ten studies, some of which have not even started recruiting participants in the current status of this dashboard. Note that the y-axis is logarithmic.

to cross this threshold that is smaller than 0.25%. In this repeated sampling the size of the study series is essentially random: we can be lucky and observe very convincing data in the early studies, making more studies superfluous, or we can be unlucky and in need of more studies. The threshold can be reached with a single study, with a two-study meta-analysis, with a three-study,.. etc, and the repeated sampling properties, like type-I error control, hold on average over all those sampling scenarios (so unconditional on the series size).

ALL-IN meta-analysis allows for meta-analyses with Type-I error control, while completely avoiding the effects of accumulation bias and multiple testing. This is possible for two reasons: (1) we do not just perform meta-analyses on study series that have reached a certain size, but continuously monitor study series irrespective of the current number of studies in the series; (2) we use likelihood ratios (and their cousins, e-values (Grünwald et al., 2019)) instead of raw Z-scores and p-values; we say more on likelihood ratios further below.

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

The ALL-IN meta-analysis in Figure 8 illustrates an improved efficiency by not setting the number of studies in advance, but let it rely on the data and be – just like the data itself – essentially random before the start of the research effort. This introduces dependencies between study results and series size that can be expressed in similar ways as *Gold Rush* accumulation bias. Yet this field of studies might make decisions differently to our *Gold Rush*: a positive nonsignificant result might not terminate the research effort, but encourage extra studies. And instead of always encouraging extra studies, a very convincing series of significant studies might conclude the research effort. If a series of studies is dependent on any such data-driven decisions, the use of conventional statistical methods is inappropriate. These dependencies actually do not have to be extreme at all: Many fields of research might be a bit like the *Gold Rush* scenario in their response to finding significant negative results of harm. A widely known study result that indicated significant harm might make it very unlikely that the series will continue to grow. So large study series will very rarely have a completely symmetric sampling distribution,

since initial studies that observe results of significant harm do not grow into large series. Hence this small aspect of accumulation bias will already invalidate conventional meta-analysis, when it assumes such symmetric distributions under the null hypothesis with equal mass on significant effects of harm and benefit.

Properties averaged over time

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

Table 1 is inspired by Senn (2014) (different question, similar answer) and represents our extreme *Gold Rush* world of study series. It takes the same approach as Figure 7 and indicates the probability to meta-analyze a one-study, two-study or three-study series of each possible form under the null hypothesis. The three study series are very biased, with two or even three out of three studies showing a positive significant effect. But the $\mathbf{P}_0$ column shows that the probability of being in this scenario is very small under the null hypothesis, as was also apparent from Figure 7. In fact, most analysis will be of the one-study kind, that hardly have any bias, and are even slightly to the left of the theoretic standard null distribution. Exactly this phenomenon balances the biased samples of series of larger size.

Table 1. Possible study series under extreme *Gold Rush* accumulation bias, with their respective probabilities $\mathbf{P}_0$ to occur under the null hypothesis. A Z-score is marked by a * and color orange (e.g. z_1^*) in case the individual study result is significant and positive ($z_1 \geq z_\alpha$ (one-sided test)) and by a $-$ (e.g. z_1^-) otherwise. The column t indicates the number of studies and the column $*$ counts the number of significant studies. The fifth and sixth column multiply $\mathbf{P}_0$ with the $*$ column and t column to arrive at an expected value $\mathbf{E}_0[*]$ and $\mathbf{E}_0[t]$ respectively in the bottom row.

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

The bottom row of Table 1 gives the expected values for the number of significant studies per series in the $* \cdot \mathbf{P}_0$ column, and the expected value for the total number of studies per series in the $t \cdot \mathbf{P}_0$ column. If we use these expressions to obtain the proportion of expected number of significant to expected total number of studies, we get the following:

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

The proportion of expected significant effects to expected series size is still α in Table 1 under extreme *Gold Rush* accumulation bias, as it would also be without accumulation bias.

This result is driven by the fact that there is a martingale process underlying this table. If a statistic is a martingale process and it has a certain value after t studies, the conditional expected value of the statistic after $t+1$ studies, given all the past data, is equal to the statistic after t studies. So if our proportion of significant positive studies is exactly α for the first study (t = 1), we expect to also observe a proportion α if we grow our series with an additional study (t = 1+1 = 2). The accumulation bias does not affect such statistics when averaged over time if martingales are involved (Doob's optional stopping theorem for martingales). You can verify this aspect by deleting the last row for $\mathbf{z_1^*}, \mathbf{z_2^*}, \mathbf{z_3^*}$ from our table and adding two rows for $t = 4$ in its place with $\mathbf{z_1^*}, \mathbf{z_2^*}, \mathbf{z_3^*}$ and either a fourth significant or a nonsignificant study. If you calculate the expected significant effects to expected series size, you will again arrive at α.

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

Multiple testing over time

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

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

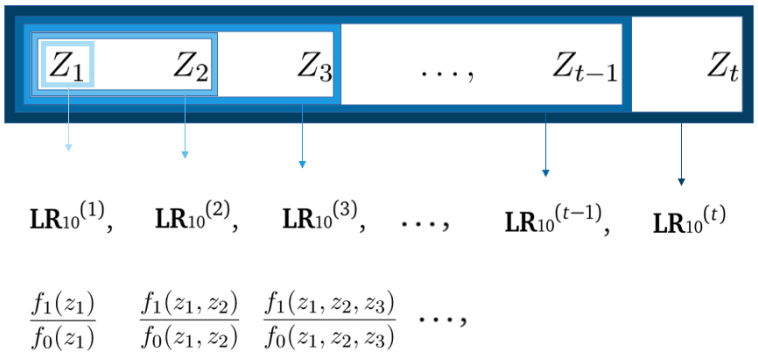

Figure 9. Likelihood Ratio martingale

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

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

The ALL-IN meta-analysis in Figure 8 in fact is based on likelihood ratios like this, and controls the type-I error by the threshold 400 at level $1/400 = 0.25\%$.

The code below illustrates that likelihood ratios can also control type-I error rates under continuous monitoring when extreme *Gold Rush* accumulation bias is at play. Within our previous simulation, we again assume a *Gold Rush* world with only true null studies and very biased two-study and three-study series. The code in Figure 11 calculates likelihood ratios for the growing study series under accumulation bias. So, here we assume that a series is analyzed for each size it reaches (so after each new study), as indicated below. Figure 11 illustrates that still very few likelihood ratios ever grow very large.

```
numSim.study <- 64000  # we're not plotting histograms, so a smaller simulation will do

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

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

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

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

dataLR.unc <- data.frame(t = c(rep(0:1, each = length(A1notA2)),
                               rep(0:2, each = length(A2notA3)),
                               rep(0:3, each = length(A3))),
                         LR = c(rep(1, times = length(A1notA2)), LR1.A1notA2,
                                rep(1, times = length(A2notA3)), LR1.A2notA3, LR2.A2notA3,
                                rep(1, times = length(A3)), LR1.A3, LR2.A3, LR3.A3),
                         series = c(rep(A1notA2, times = 2),
                                    rep(A2notA3, times = 3),
                                    rep(A3, times = 4)),
                         GoldRush = c(rep("A1notA2", times = length(A1notA2)*2),
                                      rep("A2notA3", times = length(A2notA3)*3),
                                      rep("A3", times = length(A3)*4)))
ggplot(dataLR.unc) +
  geom_line(aes(x = t, y = LR, group = series, colour = GoldRush)) +
  scale_color_discrete(name = "Gold Rush A(t)",
                       labels = c(expression(paste("A(1) = 1, A(2) = 0     ", LR[10]^(t))),
                                  expression(paste("A(2) = 1, A(3) = 0     ", LR[10]^(t))),
                                  expression(paste("A(3) = 1     ", LR[10]^(t)))))
```

Figure 10. Code to create Figure 11

If we set our type-I error rate α to 5%, and compare our likelihood ratios to $1/\alpha = 20$ we observe that less than $1/20 = 5\%$ of the study series *ever* achieves a value of $\mathbf{LR}_{10}$ larger than 20 (Figure 12). The simulated type-I error is even much smaller than 5% since in our *Gold Rush* world series stop growing at three studies, yet this procedure controls type-I error also in the case none of these series stops growing at three studies, but all continue to grow forever.

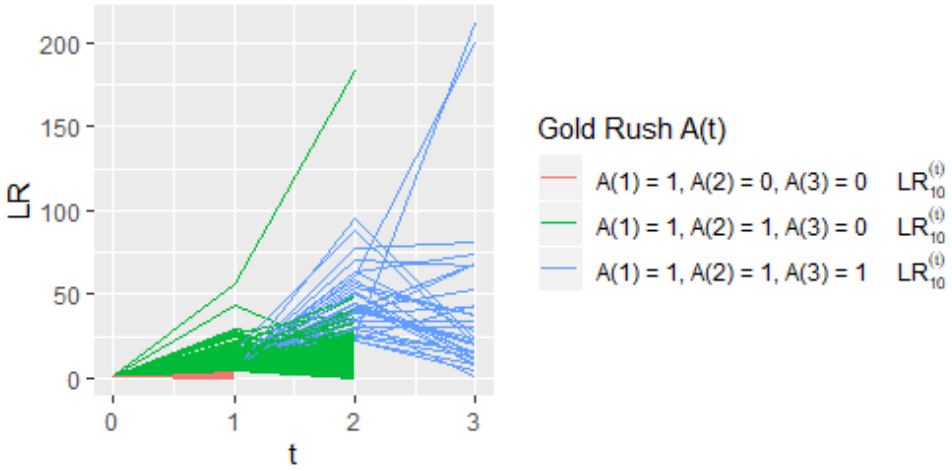

Figure 11. Unconditional sampling distributions under the null hypothesis of $\mathbf{LR}_{10}^{(t)}$ of colored one, two or three studies under extreme *Gold Rush* accumulation bias, under the assumption of equal study sample size and variance.

```
> typeIerrorLR <- mean(c(LR1.A1notA2,
+                        pmax(LR1.A2notA3, LR2.A2notA3),
+                        pmax(LR1.A3, LR2.A3, LR3.A3))
+                      > 20)
> typeIerrorLR
[1] 0.001578125
```

Figure 12. Code to calculate type-I error probability for $\mathbf{LR}_{10}^{(t)}$ averaged over series size t under extreme *Gold Rush* accumulation bias and continuous monitoring.

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

z_1^-	$A(1)=1$	$A(2)=0$	$A(3)=0$
$\mathbf{z_1^*}, z_2^-$	$A(1)=1$	$A(2)=1$	$A(3)=0$
$\mathbf{z_1^*}, \mathbf{z_2^*}, z_3^-$	$A(1)=1$	$A(2)=1$	$A(3)=1$
$\mathbf{z_1^*}, \mathbf{z_2^*}, \mathbf{z_3^*}$	$A(1)=1$	$A(2)=1$	$A(3)=1$

Note that continuous monitoring changes our *Gold Rush world* as indicated above. We can, however, also keep type-I error control if we do not continuously monitor, but only analyze the terminated series, such as in *Gold Rush world; all-series-size*, as shown in Figure 13.

```
> typeIerrorLR <- mean(c(LR1.A1notA2,
+                        LR2.A2notA3,
+                        LR3.A3)
+                      > 20)
> typeIerrorLR
[1] 0.0011875
```

Figure 13. Code to calculate type-I error probability for $\mathbf{LR}_{10}^{(t)}$ averaged over series size t under extreme *Gold Rush* accumulation bias, with analysis only at the terminated series.

The type-I error control is thus conservative, and we pay a small price in terms of power. That price is quite manageable, however, and can be tuned by setting the mean value of the alternative likelihood (arbitrarily set to `mean = 1` in the code for `calcLR` of Figure 10). More on that in Grünwald et al. (2019) and the forthcoming preprint paper on ALL-IN meta-analysis that will appear on https://projects.cwi.nl/safestats/.

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

Conclusion

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

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

Postscript

ALL-IN meta-analysis has been applied during the corona pandemic to analyze an accumulating series of studies while they were still ongoing. Each study investigated the ability of the BCG vaccine to prevent covid-19, but data on covid cases came in only slowly (fortunately). Meta-analyzing interim results and data-driven decisions improved the possibility of finding efficacy earlier in the pandemic. A webinar on the methodology underlying this meta-analysis – the specific likelihood ratios – is available on https://projects.cwi.nl/safestats/ under the name ALL-IN-META-BCG-CORONA.

Acknowledgements

My thanks go to Professor Bob Reed for inviting this contribution to his website and his patience with its publication. I also want to acknowledge Professor Peter Grünwald for checking the statistical details of this post. Daniel Lakens provided me with great advice to write this text more blog-like. Muriel Pérez helped me with the details of the martingale underlying the table.

References

Iain Chalmers and Paul Glasziou. Avoidable waste in the production and reporting of research evidence. *The Lancet*, 114(6):1341–1345, 2009.

Iain Chalmers, Michael B Bracken, Ben Djulbegovic, Silvio Garattini, Jonathan Grant, A Metin Gülmezoglu, David W Howells, John PA Ioannidis, and Sandy Oliver. How to increase value and reduce waste when research priorities are set. *The Lancet*, 383(9912):156–165, 2014.

Hans Lund, Klara Brunnhuber, Carsten Juhl, Karen Robinson, Marlies Leenaars, Bertil F Dorch, Gro Jamtvedt, Monica W Nortvedt, Robin Christensen, and Iain Chalmers. Towards evidence based research. *Bmj*, 355: i5440, 2016.

Judith ter Schure and Peter Grünwald. Accumulation Bias in meta-analysis: the need to consider time in error control [version 1; peer review: 2 approved]. *F1000Research*, 8:962, June 2019. ISSN 2046-1402. doi: 10.12688/f1000research.19375.1. URL https://f1000research.com/articles/8-962/v1.

Steven P Ellis and Jonathan W Stewart. Temporal dependence and bias in meta-analysis. *Communications in Statistics—Theory and Methods*, 38(15):2453–2462, 2009.

Elena Kulinskaya, Richard Huggins, and Samson Henry Dogo. Sequential biases in accumulating evidence. *Research synthesis methods*, 7(3):294–305, 2016.

Peter Grünwald, Rianne de Heide, and Wouter Koolen. Safe testing. *arXiv preprint arXiv:1906.07801*, 2019.

Stephen Senn. A note regarding meta-analysis of sequential trials with stopping for efficacy. *Pharmaceutical Statistics*, 13(6):371–375, 2014.

Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter Koolen. Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv preprint arXiv:2009.03167*, 2020.

Jean Ville. Etude critique de la notion de collectif. *Bull. Amer. Math. Soc*, 45(11):824, 1939.

Richard Royall. *Statistical evidence: a likelihood paradigm*, volume 71. CRC press, 1997.

A Technical Details

This blog post discusses approaches to meta-analysis that control type-I error averaged over study series size. This is called error control *surviving over time* in Ter Schure and Grünwald (2019)), as will become more clear in the technical details below.

A.1 Time: timing and chronology

Following notation from Ter Schure and Grünwald (2019), we denote the number of studies available on a certain topic by t. This number t can also indicate the *timing* of a meta-analysis, such that a meta-analysis can possibly occur at time $t = 1, 2, 3, \ldots$ up to some maximum number of studies T. The number of studies and the timing of a meta-analysis share the notion of chronology; of past, present and future studies. At $t = 3$, we have three studies available that we can possibly meta-analyse. The fact that a third study exists can depend on the result of the first and second, but can never depend on the result of a future fourth study. Analogously, our timing of a meta-analysis after three studies can depend on the results of those three studies, but never on future meta-analyses. Note that dependencies in time are *possible*, but not necessary, to apply the notation from the accumulation basis framework (Ter Schure and Grünwald, 2019). Simultaneous studies can also be described, in which case their existence cannot depend on each other. A "no dependency"-relation does require the simultaneous studies to be assigned an arbitrary chronology, but their order plays no further role than to express a set of studies as a series. In the example of this blog, however, the extreme *Gold Rush* scenario, we assume a very real chronology and deterministic dependency between all the studies in a series.

A.2 Extreme *Gold Rush* expressed in accumulation bias notation A(t)

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

$$A(t \mid z_1 \dots z_t) = \mathbf{P}[\mathcal{A}^{(t)} \mid T \geq t, z_1 \dots z_t] \cdot S(t-1 \mid z_1 \dots z_{t-1}) \tag{3}$$

In this simplified version of the *Gold Rush* scenario $S(t)$ is always either 0 or 1 if the study results z_1 and z_2 are known:

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

$S(t)$ is a survival probability, because a series can only grow to three studies (it has survive the second study ($S(2)$)) if it has also grown to two studies (it has survived the first study ($S(1)$). In contrast to the deterministic extreme *Gold Rush* in this paper, Ter Schure and Grünwald (2019) describe a probabilistic version where the probability of a replication study is larger following a significant positive result, but not always zero following a nonsignificant one. In that case we specify *hazards* of stopping after observing a certain result, which are probabilities of stopping, given that the series accumulated so far. The survival probability is defined in terms of these hazards, following standard survival analysis notation.

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

$$\begin{aligned} A(2 \mid z_1^-) &= \mathbf{P}[\mathcal{A}^{(2)} \mid T \geq 2] \cdot S(1 \mid z_1^-) = 0 \cdot 0 = 0 \\ A(2 \mid \mathbf{z_1^*}, z_2^-) &= \mathbf{P}[\mathcal{A}^{(2)} \mid T \geq 2] \cdot S(1 \mid \mathbf{z_1^*}) = 0 \cdot 1 = 0 \\ A(2 \mid \mathbf{z_1^*}, \mathbf{z_2^*}) &= \mathbf{P}[\mathcal{A}^{(2)} \mid T \geq 2] \cdot S(1 \mid \mathbf{z_1^*}) = 0 \cdot 1 = 0 \\ A(3 \mid \mathbf{z_1^*}, \mathbf{z_2^*}, z_3^- \mid T \leq 3) &= \mathbf{P}[\mathcal{A}^{(3)} \mid T \geq 3] \cdot S(2 \mid \mathbf{z_1^*}, \mathbf{z_2^*}) = 1 \cdot 1 = 1 \\ A(3 \mid \mathbf{z_1^*}, \mathbf{z_2^*}, \mathbf{z_3^*}) &= \mathbf{P}[\mathcal{A}^{(3)} \mid T \geq 3] \cdot S(2 \mid \mathbf{z_1^*}, \mathbf{z_2^*}) = 1 \cdot 1 = 1 \end{aligned}$$

A.3 Extreme *Gold Rush* conditional sampling distribution

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

Using notation from Ter Schure and Grünwald (2019) we express the accumulation of t studies as $T \geq t$, indicating that once we have t studies available, our maximum amount of studies T is at least t (it is either t or larger). $\mathcal{A}^{(t)}$ indicates the event that we perform a meta-analysis of the t studies available. We denote the conditional sampling distribution of a $z^{(t)}$-score by $f_0(z^{(t)} \mid \mathcal{A}^{(t)}, T \geq t)$, and obtain its expression by observing that $A(t)$ is a probability of a t-study meta-analysis conditioned on the data, and we need a probability of the data conditioned on the occurrence of a t-study meta-analysis; Bayes' rule transposes the conditional in the following expression:

$$\begin{aligned} f_0\left(z^{(t)} \,\middle|\, \mathcal{A}^{(t)}, T \geq t\right) &= \frac{f_0(z^{(t)}) \cdot \mathbf{P}_0\left[\mathcal{A}^{(t)}, T \geq t \,\middle|\, z^{(t)}\right]}{\mathbf{P}_0\left[\mathcal{A}^{(t)}, T \geq t\right]} \\ &= \frac{f_0(z^{(t)}) \cdot \overline{A}_0\left(t \,\middle|\, z^{(t)}\right)}{\overline{A}_0\left(t\right)}, \end{aligned} \tag{5}$$

where we define:

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

For the extreme *Gold Rush* scenario of this paper, and the sampling distribution of a three-study series illustrated in Figure 2, $\overline{A}_0\left(3\right)$ can be calculated as follows:

$$\begin{aligned} \overline{A}_0\left(3\right) &= \mathbf{E}_0[A(3 \mid Z_1, Z_2, Z_3)] \\ &= A(3 \mid \mathbf{z_1^*}, \mathbf{z_2^*}, z_3^-) \cdot \mathbf{P}_0[\mathbf{z_1^*}, \mathbf{z_2^*}, z_3^-] + A(3 \mid \mathbf{z_1^*}, \mathbf{z_2^*}, \mathbf{z_3^*}) \cdot \mathbf{P}_0[\mathbf{z_1^*}, \mathbf{z_2^*}, \mathbf{z_3^*}] \\ &= 1 \cdot \mathbf{P}_0[\mathbf{z_1^*}, \mathbf{z_2^*}, z_3^-] + 1 \cdot \mathbf{P}_0[\mathbf{z_1^*}, \mathbf{z_2^*}, \mathbf{z_3^*}] \\ &= 1 \cdot \alpha \cdot \alpha \cdot (1-\alpha) + 1 \cdot \alpha \cdot \alpha \cdot \alpha \\ &= \frac{1}{1600} \quad (\text{for } \alpha = 2.5\%) \end{aligned} \tag{6}$$

The only three-study series that have nonzero $A(t)$ are $A(3 \mid \mathbf{z_1^*}, \mathbf{z_2^*}, z_3^-)$ and $A(3 \mid \mathbf{z_1^*}, \mathbf{z_2^*}, \mathbf{z_3^*})$, such that only these have to be enumerated in expectation $\overline{A}_03)$. $\overline{A}_0\left(3 \,\middle|\, z^{(t)}\right)$ can be obtained by considering all the possible combinations of Z_1, Z_2, Z_3 that could be summarized into a specific $z^{(t)}$ and taking into account their probabilities under the null hypothesis.

The value 1/1600 explains the statement in the beginning of the code in Figure 1 that 1600 first studies are needed for each sample of a three-study series.

A.4 A(t) behaves like a survival probability

Table 2 is an extension of the table in the blogpost and shows that even though $\overline{A}_0(t)$ indicates the null hypothesis probability of accumulating t studies and meta-analyzing them, it cannot in itself tell us how often the research effort is terminated at exactly those t studies. This is caused by the fact that $A(t)$ is partly a survival probability and can be illustrated by adding a column of $\overline{A}_0(t)$ values to our table that does not add up to one.

Table 2. Possible study series under extreme *Gold Rush* accumulation bias.

	τ		$\mathbf{N^*}(\tau)$	$\overline{\mathbf{A}}_\mathbf{0}(\tau)$	$\mathbf{P}_0$	$\mathbf{N^*}(\tau) \cdot \mathbf{P}_0$	$\tau \cdot \mathbf{P}_0$
	1	z_1^-	0	1	$1-\alpha$	0	$1-\alpha$
	2	$\mathbf{z_1^*}, z_2^-$	1	α	$\alpha(1-\alpha)$	$\alpha(1-\alpha)$	$2\alpha(1-\alpha)$
	3	$\mathbf{z_1^*}, \mathbf{z_2^*}, z_3^-$	2	α^2	$\alpha^2(1-\alpha)$	$2\alpha^2(1-\alpha)$	$3\alpha^2(1-\alpha)$
$T=3$	3	$\mathbf{z_1^*}, \mathbf{z_2^*}, \mathbf{z_3^*}$	3	α^3	α^3	$3\alpha^3$	$3\alpha^3$
$\sum$					1	$\alpha+\alpha^2+\alpha^3$	$1+\alpha+\alpha^2$

A.5 The martingale underlying the table

Table 2 is slightly modified in comparison to the blog to introduce more formal notation for the *Gold Rush* stopping rule. Here we show the specific martingale underlying this table and how Doob's Optional Stopping Theorem explains the relation between the values in the bottom row of the table.

We assume that each individual study Z-score is independently sampled from a standard normal distribution with mean zero, such that the probability of obtaining a significant and positive result (if $z \geq z_\alpha$) is α. Using $\mathbf{1}(z_i)$ for the indicator function that indicates whether z_i is significant and positive, the martingale $\{M_1, M_2, M_3, \ldots\}$ underlying Table 2 is defined as follows:

$$M_t = \sum_{i=1}^{t} \mathbf{1}(z_i) - t\alpha.$$

$\{M_1, M_2, M_3, \ldots\}$ is a martingale since

$$\begin{aligned} \mathbf{E}_0\left[M_t - M_{t-1}\right] &= \mathbf{E}_0\left[\sum_{i=1}^{t} \mathbf{1}(z_i) - t\alpha - \left(\sum_{i=1}^{t-1} \mathbf{1}(z_i) - (t-1)\alpha\right)\right] \\ &= \mathbf{E}_0\left[\mathbf{1}(z_t) - \alpha\right] = \alpha - \alpha = 0. \end{aligned} \tag{7}$$

We denote the number of significant positive studies in a series of size t by $N^*(t)$ in Table 2 and express this number in terms of M_t:

$$N^*(t) = \sum_{i=1}^{t} \mathbf{1}(z_i) = M_t + t\alpha.$$

The *Gold Rush* stopping rule implies that we only stop accumulating studies at series size $t = \tau$ if we find the first nonsignificant study (z_τ^-, where $\tau = \min_t\{\mathbf{1}(z_t) = 0\}$) or if we arrive at the maximum series size $t = T$. This stopping rule forces us to stop accumulating studies at either series size τ or at size T, whichever comes first. So we stop at $\tau \wedge T$. We express the expected number of studies that is significant and positive in terms of the expectation of the martingale under this stopping rule:

$$\mathbf{E}_0[N^*(\tau \wedge T)] = \mathbf{E}_0[M_{\tau \wedge T}] - \mathbf{E}_0[\tau \wedge T]\alpha,$$

and since $\tau \wedge T$ is always finite, by Doob's Optional Stopping theorem we have:

$$\begin{aligned}
\mathbf{E}_0[M_{\tau\wedge T}] = \mathbf{E}_0[M_1] = \mathbf{E}_0\left[\sum_{i=1}^{1} \mathbf{1}(z_i) - 1\alpha\right] = \mathbf{E}_0[\mathbf{1}(z_i)] - \alpha = \alpha - \alpha = 0 \\
\text{such that} \\
\mathbf{E}_0[N^*(\tau \wedge T)] = \mathbf{E}_0[\tau \wedge T]\alpha \\
\text{and} \\
\frac{\mathbf{E}_0[N^*(\tau \wedge T)]}{\mathbf{E}_0[\tau \wedge T]} = \alpha.
\end{aligned} \tag{8}$$

This is shown for the *Gold Rush* stopping rule and $T = 3$ by Table 2, but holds for any stopping rule and finite T.