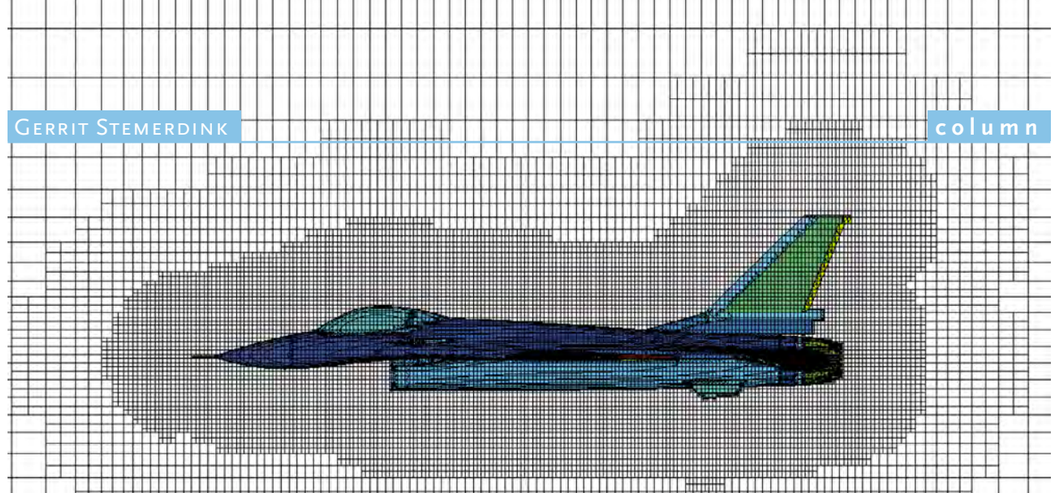
Gerrit Stemerdink

column



Bezuinigen op verkiezingen met slimme wiskunde

Slim gebruik van wiskunde kan tot aanzienlijke besparingen leiden. De vele toepassingen van OR zijn hier een voorbeeld van.

Soms zitten die besparingen op plekken die wat minder bekend zijn. Zo kent de numerieke wiskunde de methode van het Adaptive Grid. Bij het berekenen van bijvoorbeeld de luchtstromingen rond een vliegtuig wordt een raster van gelijke vierkanten over het toestel gelegd en wordt voor iedere cel de in- en uitstroom berekend. Het verschil tussen die twee resulteert dan in een over- of onderdruk etc. Die berekening wordt voor iedere cel vele malen herhaald, telkens een tijdstapje verder. Steeds wordt gekeken naar de druk in de naastliggende cellen, immers die bepaalt de mate van in- en uitstroom. Dit kost enorm veel rekentijd, zeker als de cellen en de tijdstappen klein zijn. Bij een Adaptive Grid worden de cellen eerst vrij groot gekozen, vervolgens wordt gekeken in welke cellen zich tijdens de eerste tijdstappen grote veranderingen voordoen. Alleen díe cellen worden vervolgens verkleind, de rest wordt ongemoeid gelaten. Dat proces wordt vele malen herhaald: de cellen worden kleiner naarmate er meer verandering in plaats vindt, vandaar de naam Adaptive Grid. Het idee hierachter is dat men geen rekentijd hoeft te verspillen aan die delen van het raster waar de veranderingen klein zijn, men zet de computerkracht hoofdzakelijk daar in waar 'het ertoe doet'.

Toen ik weer eens veel te lang en veel te laat naar CNN keek tijdens de dagen direct na de Amerikaanse verkiezingen van 2020 bedacht ik dat men hier eigenlijk prima dat principe van het Adaptive Grid zou kunnen toepassen. De uitslag wordt, door het in mijn ogen vreemde systeem van 'the winner takes all' kiesmannen, in de praktijk slechts bepaald door de uitkomsten zo'n 8 tot 10 staten. De toewijzing van de kiesmannen aan Democraten of Republieken in de overige staten ligt immers al van tevoren vast. Een Democraat in Texas kan stemmen tot hij een ons weegt, de kiesmannen van zijn staat gaan toch onveranderlijk naar de Republikeinen. Waarom zou men dan in zulke staten nog iedere vier jaar verkiezingen organiseren, dat is verspilde tijd, geld en moeite. Als er in een staat de uitslag een verschil van meer dan bijvoorbeeld 5% tussen de beide kampen laat zien wordt die staat bij de volgende verkiezing overgeslagen. De vierjaarlijkse verkiezingen worden dan alleen in die staten gehouden die in de praktijk de einduitslag bepalen, zo kan men tot aanzienlijke besparingen komen. Ook hier weer alleen je energie gebruiken daar waar 'het ertoe doet'.

Denk overigens niet dat dit idee van mij revolutionair is. In 1958 las ik een science fiction verhaal ('Franchise' door Isaac Asimov) over de presidentsverkiezing in 2008, toen nog een verre toekomst. Daar gaat het nog veel verder. Op de dag van de verkiezing wordt slechts één enkele kiezer met veel ceremonieel van huis gehaald en naar het stemlokaal gebracht. Daar brengt hij zijn stem uit en dat is het. Amerika was een Electronic Democracy geworden en de computer Multivac, die alles van alle inwoners wist, had bepaald dat deze ene kiezer het perfecte gemiddelde is van alle kiezers in het land: hij vertegenwoordigt daarom in zijn eentje alle kiezers. Dat zou pas een grote besparing zijn! Maar of we zoiets zouden moeten willen?

Gerrit Stemerdink is eindredacteur van *STAtOR*.
E-mail: gjstemerdink@hotmail.com

ACCUMULATION BIAS

How to handle it ALL-IN

Judith ter Schure

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

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. Here, we discuss 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 assume that we always stop performing studies when we've reached three, so we could possibly perform a meta-analysis that includes $t = 1$, 2 or 3 studies: t indicates the number of studies as well as the timing of the meta-analysis. We summarize the results of individual studies in a single per-study Z-score (z_1 for the first study, z_2 for the second, etc). A significant positive study is shown as z_i^* ($z_i > z_\alpha$ with $z_\alpha = 1.96$ for $\alpha = 2.5\%$) and a nonsignificant or negative one as z_i^-. Our *Gold Rush* world consists of the possible study series in the box below.

Here A(t) denotes whether we accumulate **and** analyze the 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. $z^{(3)}$ indicates the Z-score of a fixed or common effect meta-analysis The effects of accumulation bias are not limited to fixed-effects meta-analysis (see for example Kulinskaya et al. (2016)), but we use fixed-effects meta-analysis as a simple illustration.

We observe in our *Gold Rush* world below 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.

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 R Code 1. A fixed-effect meta-analysis Z-score weights the study estimates by their precision (inverse variance). Here we assume equal variance and large enough (and equal) sample size to estimate the variance without error, in which case the fixed-effects $z^{(t)}$-score reduces to 1/sqrt(t)*sum(zs).

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 R Code 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 capture 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.

Figure 1 plots 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.

We observe in Figure 1 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

Gold Rush world

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

```r
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])))
```

R Code 1 to simulate Figure 1

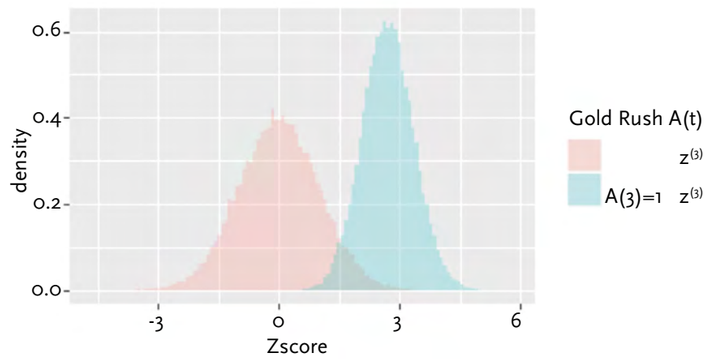


Figure 1. Sampling distributions under the null hypothesis of fixed-effects meta-analysis Z-scores $Z^{(3)}$ of three studies with and without extreme Gold Rush accumulation bias A(t), under the assumption of equally large study sample size and equal variance

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 1 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.

Accumulation bias can be efficient

The steps in the code from R Code 1 that arrive at the sampling distributions in Figure 1 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 the next section we take the unconditional view.

ALL-IN meta-analysis

Figure 2 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. This is important to note that such data-driven decisions would invalidate conventional meta-analysis by introducing accumulation bias.

To interpret Figure 2, 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

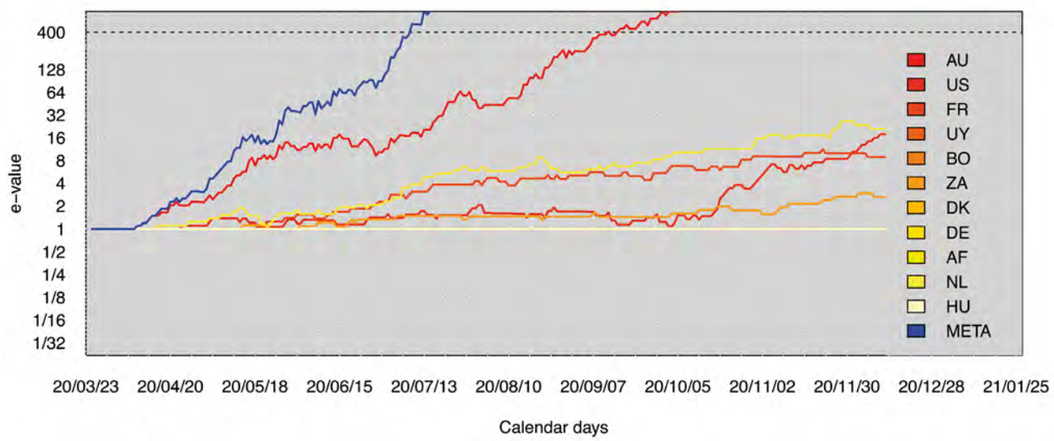


Figure 2. Dashboard of an ALL-IN meta-analysis of between one and eleven studies (with fake data), some of which have not even started recruiting participants in the current status of this dashboard. Note that the y-axis is logarithmic

analysis, however, does not matter for its evidential value. Some studies (like the Australian one) are much larger than others, such that under a lucky scenario this study could reach the evidential threshold even before other studies start observing data. This threshold (indicated at 400) controls type-I errors at a rate of $\alpha = 1/400 = 0.0025$ (details in the final section). So in repeated sampling under the null, the combined studies will only have a probability to cross this threshold that is smaller than 0.25%. In this repeated sampling the size of the study series is essentially random: we can be lucky and observe very convincing data in the early studies, making more studies superfluous, or we can be unlucky and in need of more studies. The threshold can be reached with a single study, with a two-study meta-analysis, with a three-study ... etc, and the repeated sampling properties, like type-I error control, hold on average over all those sampling scenarios (so unconditional on the series size).

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

Properties averaged over time

Table 1 is inspired by Senn (2014) (different question, similar answer) and represents our extreme *Gold Rush* world of study series. The three study series are very biased, with two or even three out of three studies showing a positive significant effect. But the $\mathbf{P}_\circ$ column shows that the probability of being in this scenario is very small under the null hypothesis. 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.

The bottom row of Table 1 gives the expected values for the number of significant studies per series in the $* \cdot \mathbf{P}_\circ$ column, and the expected value for the total number of studies per series in the $t \cdot \mathbf{P}_\circ$ 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:

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

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. A martingale for a series of studies can be thought of as a sequence of statistics, updated after each study, for which the following holds: if the statistic has a certain value after t studies, the conditional expected value of the statistic when the next study is added, so conditional on what is known so far, is equal to the statistic after t studies. The accumulation bias does not affect such statistics when averaged over time (Doob's Optional Stopping Theorem for martingales). You can get a sense of this theorem for Gold Rush accumulation bias by deleting the last row for z_1^*, z_2^*, z_3^* from our table and adding two rows for $t = 4$ in its place with z_1^*, z_2^*, 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.

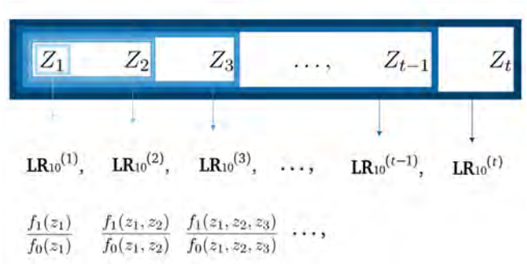
Table 1. Possible study series under extreme Gold Rush accumulation bias, with their respective number of significant studies (*) and probabilities (Po) to occur under the null hypothesis

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	$\alpha + \alpha^2 + \alpha^3$	$1 + \alpha + \alpha^2$

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) *any-time valid* (Ramdas et al., 2020). Their methods rely on martingales that are nonnegative (Ramdas et al., 2020), specifically the likelihood ratio. For a meta-analysis *Z*-score, a martingale process of likelihood ratios could look as follows:



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:

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

The ALL-IN meta-analysis in Figure 2 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 R-Code 2 illustrates that likelihood ratios can also control type-I error rates under continuous monitoring when extreme *Gold Rush* accumulation bias is at play. Just as in 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 R Code 2 calculates likelihood ratios for the growing study series under accumulation bias. Figure 3 illustrates that still very few likelihood ratio processes ever grow very large.

If we set our type-I error rate α to 5%, and compare our likelihood ratios to $1/\alpha = 20$ we observe that less than 1/20 = 5% of the study series *ever* achieves a value of **LR**10 larger than 20 (R Code 3). 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.

```
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])))
```

R Code 2 to create Figure 3

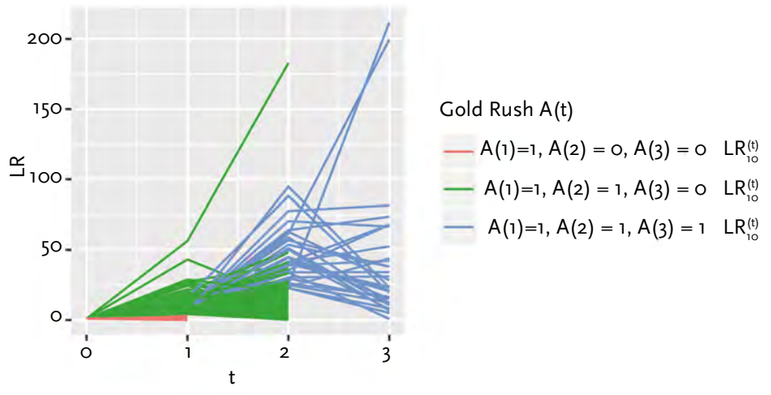


Figure 3. Samples under the null hypothesis of $LR_{10}^{(t)}$ of one, two or three studies under extreme Gold Rush accumulation bias, under the assumption of equally large study sample size and equal variance.

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 R Code 2). More on that in Grünwald et al. (2019).

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.

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.

STAtOR 2020-4 contained an article 'Nieuwe statistiek voegt wereldwijd corona-onderzoek samen' that explained the ALL-IN approach in terms of e-values and gambling. This interpretation is closely related to the notion of martingales as a *fair game*. Likelihood ratios are e-values and e-values are (super)martingales and can therefore be interpreted as the betting profit of a fair game.

References

The list of references, technical details (e.g. the specific martingale underlying the table) and a more elaborate version of this article are available on the VVSOR-websie. The digital article also links to the full R code that runs the simulations and produces the plots.

Judith ter Schure's research interests lie in foundations of statistics as well as in applied work. She divides her time between her PhD research on ALL-IN meta-analysis at CWI, freelance statistical consultancy (significanthelp.nl) and board membership (treasurer) of VVSOR.
E-mail: mail@judithterschure.nl

```
> typeIerrorLR <- mean(c(LR1.A1notA2,
+                        pmax(LR1.A2notA3, LR2.A2notA3),
+                        pmax(LR1.A3, LR2.A3, LR3.A3))
+                        > 20)
> typeIerrorLR
[1]  0.001390625
```

R Code 3 to calculate type-I error probability for continuous testing $LR_{10}^{(t)}$ averaged over series size t under extreme Gold Rush accumulation bias