Addressing exaggeration of effects from single RCTs

Randomised controlled trials are often presented as the gold standard for testing new medical treatments. In the early stages of research, however, reports from single trials are likely to show exaggerated effect estimates. **Erik van Zwet, Simon Schwab** and **Sander Greenland** explain why – and propose a remedy

Nearly three quarters of a century ago, the UK Medical Research Council published a study of the effect of streptomycin on tuberculosis. This landmark study has been cited as the first modern randomised controlled trial (RCT). Today, RCTs are often presented as the gold standard to assess the efficacy of an intervention or treatment, where efficacy is defined as the effect under ideal conditions. Although practitioners want to know treatment *effectiveness* (effects in the less-than-ideal conditions of medical practice), large RCTs such as the RECOVERY trial have been at the centre of the global search for effective treatments for Covid-19.

A key part of designing any RCT is to estimate how large the trial's sample size needs to be. Too small and it risks failing to provide convincing evidence regarding effects; too large and it becomes infeasibly expensive. The usual approach is to ensure that there is at least 80% or 90% probability or "power" for obtaining $p < 0.05$ (and thus "detecting" an effect at the 0.05 level) under the assumption that the effect of the treatment is a particular size. The European

The deconvolution trick

We can estimate the *joint* distribution of *z* and the SNR from a sample of pairs (*b*, *s*) as follows. We start by modelling the marginal distribution of *z* as a mixture of zero-mean normal distributions, which we can estimate directly from the observed *z*-values. For simplicity we will not distinguish the direction of the effects. Hence, we judge the fit of our mixture to the histogram symmetrised around zero, which is shown in the bottom panel of Figure 1. In doing so we are assuming that true effects are symmetrically distributed about zero. To drop the mean-zero assumption we could symmetrise around the empirical mean instead, but in our example it makes little difference. We could also drop the symmetry step by fitting a distribution with a skew parameter.

Given *b* is normally distributed with mean β and standard deviation *s*, it follows that *z* is the sum of the SNR and independent standard normal noise. This means that the marginal distribution of *z* is the *convolution* of the distribution of the SNR and the standard normal density. Therefore, we can obtain the marginal distribution of the SNR by *deconvolution*. Since the distribution of *z* is a mixture of normal distributions, this is easy; just subtract 1 from the variances of the components! The dashed curve in Figure 1 shows the result.

The actual power is a transformation of the SNR:

$$\text{power(SNR)} = \Phi(-1.96 - \text{SNR}) + 1 - \Phi(1.96 - \text{SNR})$$

where Φ denotes the cumulative distribution function of the standard normal distribution. We generated a sample of size 1 million from the estimated distribution of the SNR and applied the transformation. We show the resulting histogram in Figure 2.

We already know the conditional distribution of *z* given the SNR, and combining that with the marginal distribution of the SNR gives us the joint distribution of *z* and the SNR. For the mathematical details, we refer to the appendix of van Zwet *et al*.[3]

Medicines Agency's ICH E9 guideline on the statistical principles for clinical trials states: "The treatment difference to be detected may be based on a judgement concerning the minimal effect which has clinical relevance in the management of patients or on a judgement concerning the anticipated effect of the new treatment, where this is larger."

While a trial may be designed to have 80% power to detect a particular effect of clinical relevance, that does not mean it has 80% probability of yielding $p < 0.05$: the latter probability depends on the actual effect of the treatment, so we refer to it as the *actual* power.

Of course, the actual power cannot be observed directly, but we can estimate its distribution among a collection of RCTs. We have analysed the primary results of 23,551 RCTs of treatment efficacy that are available in the Cochrane Database of Systematic Reviews (CDSR).[1] The Cochrane Collaboration is a global independent network that aims to gather and summarise the best evidence from medical research. It attempts to collect all available studies on a particular topic – both published and unpublished. While there is evidence that the database may still suffer from some publication bias and dubious research practices,[2] the CDSR is currently the largest and most comprehensive collection of evidence on medical interventions.

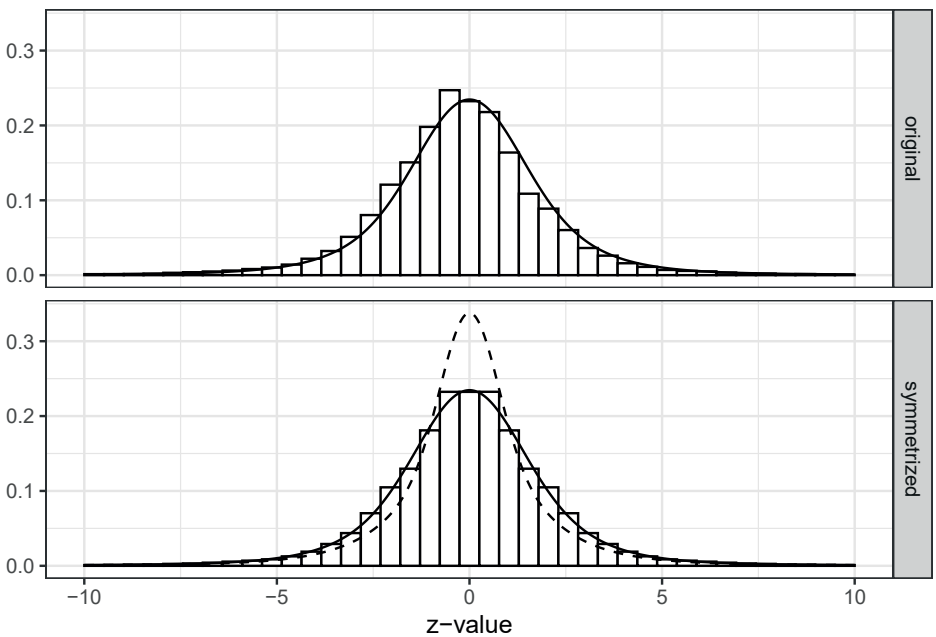

Figure 1: (Top) The histogram of the observed *z*-values, together with our fit based on a mixture of 4 zero-mean normal distributions. (Bottom) The symmetrised histogram together with the same fit (solid line) and the resulting fit for the SNR (dashed line). To symmetrise the histogram we use the R command `hist(c(-z,z))`.

Assessing the actual power of RCTs

We can capture the essence of an ideal RCT as a set of three numbers, to which we assign the notation β, *b*, and *s*. Here β represents the true effect under investigation. For example, patients treated with a particular drug might have a lower risk of mortality than those given a placebo. For statistical purposes, such an effect is often expressed as the logarithm of a hazard-rate ratio or a similar measure. What the RCT gives us, however, is just *b*, an estimate of this effect, together with its standard error *s*.

If the trial was not too small and was conducted and analysed properly, then it is

Erik van Zwet is an associate professor of medical statistics at the Leiden University Medical Center.

Simon Schwab is a postdoctoral researcher at the Center for Reproducible Science and the Department of Biostatistics at the University of Zurich.

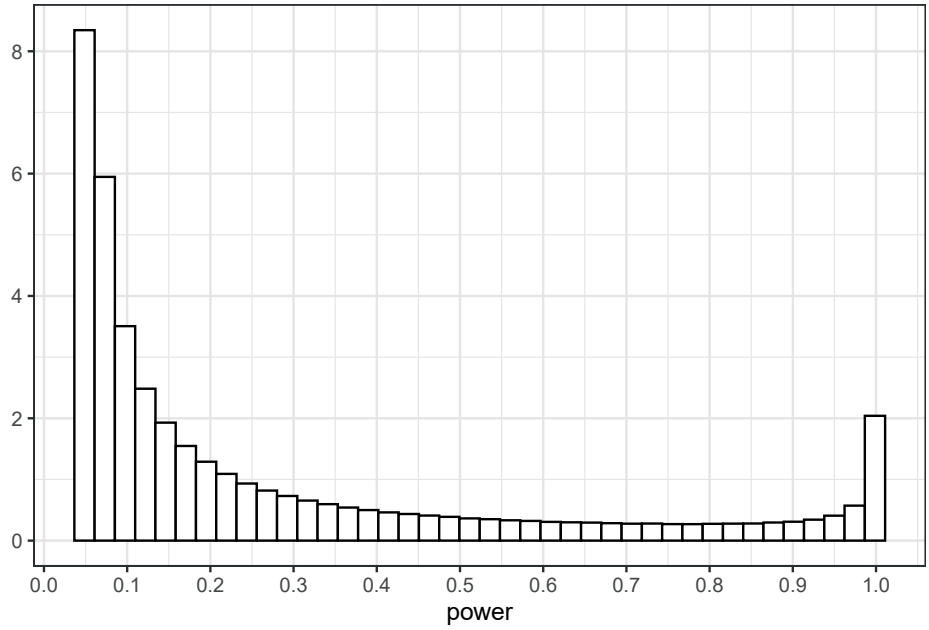

Figure 2: Histogram of a sample of size 1 million from the estimated distribution of the actual power among the RCTs in the Cochrane database.

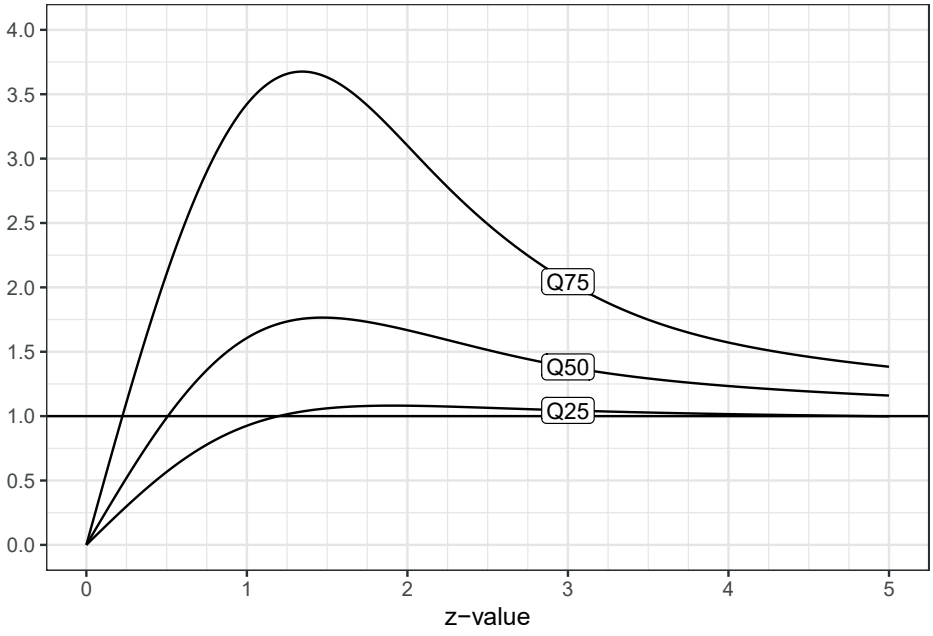

Figure 3: The distribution of the exaggeration ratio $R = |b/\beta| = |z/\text{SNR}|$ conditional on the z-value, represented by its three quartiles.

reasonable to assume that b is approximately unbiased and normally distributed. Moreover, we can then ignore the uncertainty in the estimated standard error. We define two new variables: $z = b/s$ and the signal-to-noise ratio, $\text{SNR} = \beta/s$.

We show the histogram of the z-values in Figure 1. While the distribution may appear Gaussian, it has heavier tails. We also note that 7% of the observed z-values exceed 4 in absolute value. We believe that scepticism is in order with such large values as they may well have been distorted by systematic errors or even fraud.

With the z-value and the SNR, we can apply a mathematical trick to give us insight into the distribution of the actual power of an RCT given the observed z-value; see "The deconvolution trick" (page 17).

The outcome – shown graphically in Figure 2 – is disturbing. It shows that around nine out of ten RCTs have actual power less than 80%. The mean actual power is just 28%, but the distribution is so skewed towards low values that the median actual power is more relevant, and more shocking: just 13%. In other words, most RCTs are radically underpowered to detect the *true* effect.

The extreme uptick on the right-hand side of Figure 2 is a reflection of the large z-values we already noticed in Figure 1. Discounting those as "outliers" would have resulted in a more sceptical distribution.

Although most trials do not have anything near their targeted 80% or 90% power, this does not necessarily mean that their sample size calculations are mistaken mathematically. It only indicates that most treatments do not have the effect that was used to design the trial. Finding good treatments is not easy! But as we shall see, the lack of actual power of most RCTs has serious implications for how we should interpret their outcomes.

Low actual power and the winner's curse

On the face of it, a lack of adequate actual power simply means that an RCT is too small to provide convincing evidence for the effect that is really there, and thus prone to miss important effects – a problem long noted for typical RCTs.[4] But there is a flip side to this. If an RCT *does* produce such evidence, the suspicion is that the estimated effect size must be exaggerated – and thus is unlikely to be replicated in later studies.

This exaggeration is sometimes captured via the expectation of the ratio $R = |b/\beta|$ given that the p-value fell below 0.05, that is, $E(R|\beta, s, |z| > 1.96)$, the "exaggeration ratio", which is typically much greater than 1. This is a version of the so-called "winner's curse" because obtaining a "statistically significant" result could be considered a win,

Sander Greenland is emeritus professor of epidemiology and statistics at University of California, Los Angeles.

The exaggeration ratio increases with decreasing actual power – so the lower the power, the greater the effect of the winner's curse

but as a consequence one is cursed with an overestimate of the true effect. This is often cited as a leading cause of replication failure and is a manifestation of the well-known phenomenon of regression towards the mean, in which outlying initial measurements tend to shrink upon replication towards the average value in their parent population.[5]

In our example β is the log of a ratio measure; thus it may be more interpretable to use the exaggeration ratio on the original hazard scale, which reduces to $\exp(|b - \beta|)$, but for mathematical simplicity we instead use $|b/\beta|$, ignoring the possibility that b and β have different signs.

The exaggeration ratio increases with decreasing actual power, so that the lower the power, the greater the effect of the winner's curse. This is very bad news, given our finding that most RCTs have such low actual power. Deleting outliers with very large apparent z-values would have increased our estimated exaggeration ratios beyond the already profound levels seen here.

We can quantify the problem of exaggeration among the RCTs in the Cochrane database as follows. Dividing the numerator and denominator by the standard error, we see that $R = |b/\beta| = |z/\text{SNR}|$. Since we have the joint distribution of the z-value and the SNR (see "The deconvolution trick"), we can compute the distribution of R given the observed z-value. In Figure 3 we show the three quartiles of this conditional distribution.

The quartiles show that if we observe $z = 1.96$, a finding that is borderline "statistically significant" ($p = 0.05$), there is a 75% chance that the effect will be overestimated (Q25), a 50% chance that the effect is overestimated by at least a factor of 1.7 (Q50), and a 25% chance that the effect is overestimated by more than a factor of 3 (Q75). This is a truly alarming degree of exaggeration.

Figure 3 shows the conditional distribution of the exaggeration given the z-value across all the efficacy RCTs in the Cochrane database. Of course, this distribution is not the same in different subsets of RCTs. In Figure 4 we stratified the RCTs on having smaller or larger standard error than the median. The trials with the largest standard errors tend to have particularly low actual power, and hence the exaggeration is even more severe than in Figure 3. For the trials with the smallest standard errors, it is the other way around.

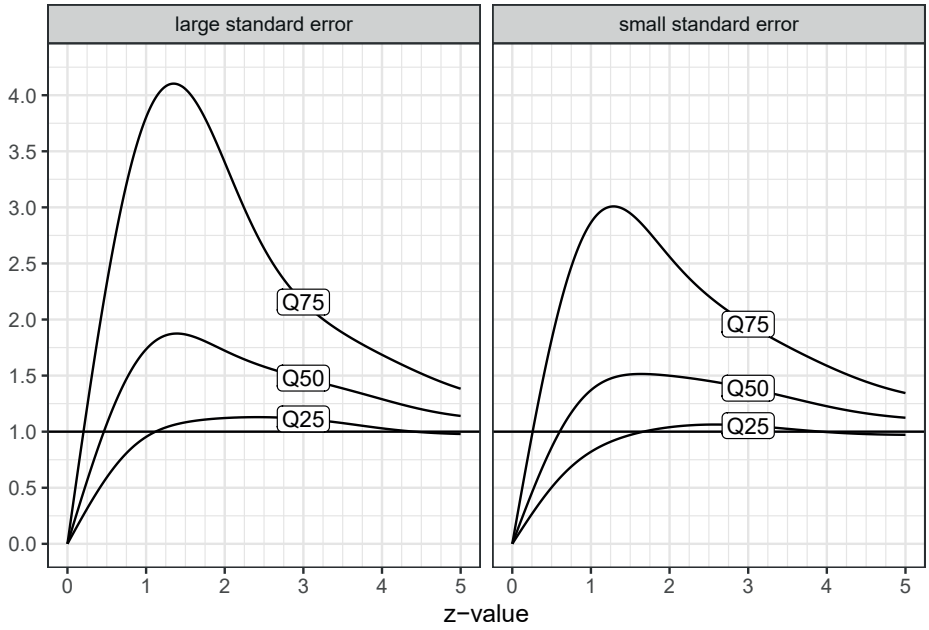

Figure 4: The distribution of the exaggeration ratio $R = |b/\beta| = |z/\text{SNR}|$ conditional on the z-value, represented by its three quartiles. We distinguish RCTs with large and small standard errors.

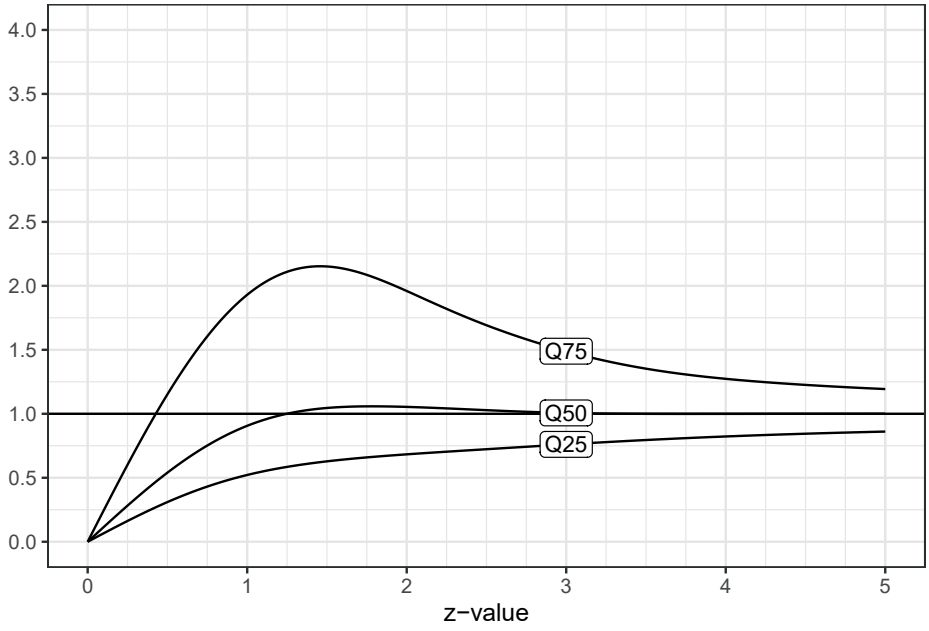

Figure 5: The distribution of the exaggeration ratio $|B/\beta|$ for the shrinkage estimator B, showing a greatly reduced risk of exaggeration.

The ANDROMEDA-SHOCK trial

The aim of the ANDROMEDA-SHOCK trial was to determine if treating patients with septic shock on the basis of capillary refill time (CRT) is superior to treating them on the basis of blood lactate levels.[6] The primary outcome was survival at 28 days and the primary effect parameter was the hazard ratio (HR) which was estimated at 0.75 with 95% compatibility ("confidence") interval of (0.55, 1.02). This result is most compatible with a 25% mortality-rate reduction, but the interval included the null and the investigators mistakenly stated that the experimental strategy "did not reduce all-cause 28-day mortality".

By applying the shrinkage estimator, we can still interpret the results of this particular RCT in light of what we know about RCTs in general. This means that we will view the ANDROMEDA-SHOCK trial as a "typical" RCT in the sense that its SNR is exchangeable with that of the other RCTs in the Cochrane database. That is, we use *only* the information that the ANDROMEDA-SHOCK trial is an RCT, and none of the particular features of the trial such as the disease, the treatment, the population, the sample size, and the type of outcome.

Correcting for the exaggeration effect, we find that the central estimate shrinks to a less impressive – but still potentially useful – HR of 0.84. The 95% interval also changes to (0.62, 1.07), which is both shifted and narrower.

We can also compute the conditional probability that the true hazard ratio is less than 1, given the observed *z*-value. It turns out that even after correction, the probability that CRT is the better approach is 91%, which is hard to square with the original report's pessimistic conclusion.

We have made an app to facilitate these calculations: vanzwet.shinyapps.io/shrinkrct.

Countering the winner's curse

RCTs often have low actual power, which implies that when we observe a z-value just a little larger than 1, there is already more than a 75% chance that the estimated effect is exaggerated. Fortunately, we can use the information in the Cochrane database to reduce this exaggeration, under the assumption that the trial we are examining is exchangeable with those in the database.

As explained in "The deconvolution trick", we can estimate the joint distribution of the z-value and the SNR among the efficacy RCTs in the Cochrane database. From this joint distribution, we can obtain the conditional expectation of the SNR given the z-value. Since SNR = β/s, this suggests an alternative estimator for β, namely $B = s \times E(\text{SNR}|z)$. The details are somewhat technical, but B is not difficult to compute; see van Zwet *et al*. for a few lines of R code.[3]

B is a "shrinkage" estimator because it can be shown that its magnitude is always less than that of the unbiased estimator b. The appropriate amount of shrinkage is estimated from the distribution of the z-values in the Cochrane database. This is equivalent to using information about the distribution of the actual power. In Figure 5 we can see that the shrinkage almost completely remedies the exaggeration seen in Figure 3.

It is also possible to obtain a 95% interval for the true effect, β. We simply multiply the 2.5th and 97.5th percentiles of the conditional distribution of the SNR given z by the standard error s. We demonstrate this calculation for a real-life RCT in "The ANDROMEDA-SHOCK trial". The resulting 95% interval is only empirical Bayesian in that it does *not* guarantee 95% coverage regardless of (or conditional on) the true effect size in the study under analysis. Instead, it provides approximate coverage unconditionally over random selection of studies from the shrinkage (prior) distribution we estimated from the Cochrane database, with lower coverage due to underestimation for studies of truly large effects, balanced by higher coverage for studies of effects near zero.

In essence, then, we are compensating for the potential exaggeration in any given RCT by applying a correction based on a property extracted from a large sample of RCTs. Technically, this is made possible by conditioning on the z-value. But this is only valid if we regard any given RCT as representative of ("exchangeable with") the RCTs in the database. This means that we are effectively ignoring any peculiarities of the RCT under study such as its sample size, type of outcome, type of control, treatment, inclusion/exclusion criteria, etc. Such features can make a big difference, as we already saw in Figure 4 where we stratified on the standard error. We can take the characteristics of the RCT under study into account by performing a bespoke Bayesian analysis. The global approach we present here can still serve as an initial sceptical screen of results, especially for readers without access to full details of a study (including its data).

We are compensating for potential exaggeration by applying a correction based on a property extracted from a large sample of RCTs

Discussion

We want to stress that decisions should rarely be made solely on the basis of a single RCT. The Cochrane Collaboration enterprise stems from the worthy goal of evidence aggregation (including meta-analysis) to interpret the findings of multiple trials (see "How to view 'underpowered' studies"). For that purpose, it is essential that each study presents its original point estimate and standard error (b, s) or interval estimate to enter into the meta-analysis, as well as enough detail about its design, conduct and analysis to judge whether those statistics can be trusted. Shrinkage estimates and related Bayesian results assume the trial has passed this preliminary screen to provide "best bets" about the effect in the study based on the study data, tempered by what previous studies have seen (including, especially, observations that later estimates tend to regress towards the null). And estimation of shrinkage from a repository of studies assumes that the repository is not afflicted by publication bias – an assumption that is unlikely to be exactly true even for the Cochrane database.[2]

If an RCT has been conducted and reported competently (a big "if", to be sure)

then it yields an unbiased estimate of the treatment effect along with a valid 95% compatibility ("confidence") interval. But these unbiasedness and validity properties refer to mean values and interval coverage over hypothetical trial repetitions, and thus are of unclear relevance after the data are in. Once the estimate *b* and its standard error *s* have become available, we need to consider properties that take account of these observations. Conditionally on the observed *z*-value, the estimate is certainly not "unbiased" in any sense of the word, as can be seen in Figure 3. When the *z*-value is larger than 1, it is likely that *b* overestimates the true effect. This large exaggeration is a direct consequence of the fact that most RCTs have low SNRs, which also leads to low actual power, that is, low probability of getting $p \leq 0.05$ under the *true* effect.

The fact that RCTs tend to have low SNRs and hence low actual power (Figures 1 and 2) is presumably due to various factors such as budgetary and logistic constraints, naïve optimism about what the true effect may be, and the fact that it is very difficult to come up with new treatments that have a large benefit over existing best practices.

Thus, despite pleas for larger sample sizes, we doubt there will be substantial increases in SNRs or actual power of RCTs in the foreseeable future. We have proposed a simple way to counter the resulting exaggeration of effect estimates by taking into account information about the distribution of SNRs among RCTs. The resulting shrinkage estimator largely remedies the problem, as can be seen in Figure 5. The estimator is easy to compute from the usual effect estimate and its standard error, which are routinely reported. In using it we should not forget the fundamental law of sound statistics: the "no free lunch principle". The cost of shrinkage is its reliance on exchangeability assumptions, which can and should be examined by probing study details that might call those into question.

Our approach is reminiscent of the Bayesian procedure of combining the outcome of a study with a "prior" capturing previous insight and experience. We do expect that many Bayesians would want to use more information than the mere fact that we are looking at an RCT which resembles a typical study in the Cochrane database; but they might still take our estimate as a reference point for further analyses. On the other hand, many frequentists may prefer to not use any external information at all. We hope that both sides will see merit in using our method alongside conventional procedures, especially when other forms of prior information are controversial, complex, or unavailable. ■

How to view "underpowered" studies

Single trials can be highly misleading about treatment efficacy in either direction simply from the play of chance, especially when the conclusions pivot on whether or not the *p*-value crosses a conventional "significance" cut point. Yet even noisy "underpowered" trials can contribute valuable information if they are well conducted and interpreted not to force conclusions or decisions, but instead as but one contribution to a collective effort (which is what the Cochrane Collaboration is about). To see this point, consider that the result of a multicentre trial is a pooled analysis of many trials (one from each centre), each far too small ever to be taken seriously on its own, yet designed and conducted under a uniform protocol to ensure validity and to make pooling seamless.

What matters most is whether the information from the study is valid within its statistical limits, which can only be ensured through proper design and execution of patient recruitment, treatment assignment, and follow-up. Size and other power-related considerations are accounted for by the span of the interval estimate, and thus proper interpretation should be built around the full range of the interval (not just whether it includes the null). With this view, the problem of being "underpowered" is less a problem of the trial than a problem of misreporting "no effect" just because the *p*-value happens to fall above 0.05 or the 95% interval includes the null[7] – as it likely will if the power is under 50%.

This problem arises often in research on secondary endpoints such as safety outcomes, in which every single trial may be very underpowered (unlikely to get $p < 0.05$) even if they would clearly signal an effect if combined properly. By focusing on "significance" or power of individual studies, the entire body of trials may instead be misinterpreted as having found nothing. This mistake goes beyond the usual one of confusing absence of evidence with evidence of absence, since now there *is* evidence of presence to which observers are blind due to failure to properly merge that evidence across studies.

Disclosure statement
The authors declare no conflicts of interest.

References

1. Schwab, S. (2020) Re-estimating 400,000 treatment effects from intervention studies in the Cochrane Database of Systematic Reviews [Data set]. Open Science Framework. doi.org/10.17605/OSF.IO/XJV9G.
2. Kicinski, M., Springate, D. A. and Kontopantelis, E. (2015) Publication bias in meta-analyses from the Cochrane Database of Systematic Reviews. *Statistics in Medicine*, **34**(20), 2781–2793.
3. van Zwet, E. W., Schwab, S. and Senn, S. J. (2021) The statistical properties of RCTs and a proposal for shrinkage. *Statistics in Medicine*. doi.org/10.1002/sim.9173
4. Freiman, J. A., Chalmers, T. C., Smith, H. Jr. and Kuebler, R. R. (1978) The importance of beta, the Type II error and sample size in the design and interpretation of the randomized control trial: Survey of 71 "negative" trials. *New England Journal of Medicine*, **299**, 690–694.
5. Ioannidis, J. P. (2008) Why most discovered true associations are inflated. *Epidemiology*, **19**(5), 640–648.
6. Hernández, G., Ospina-Tascón, G. A., Damiani, L. P., *et al*. (2019) Effect of a resuscitation strategy targeting peripheral perfusion status vs serum lactate levels on 28-day mortality among patients with septic shock: The ANDROMEDA-SHOCK randomized clinical trial. *Journal of the American Medical Association*, **321**(7), 654–664.
7. Greenland, S. (2017) The need for cognitive science in methodology. *American Journal of Epidemiology*, **186**, 639–45

We should not forget the fundamental law of sound statistics: the "no free lunch principle"