The statistical properties of RCTs and a proposal for shrinkage

Zwet, E. van; Schwab, S.; Senn, S.

Note: To cite this publication please use the final published version (if applicable).

RESEARCH ARTICLE

Statistics in Medicine WILEY

The statistical properties of RCTs and a proposal for shrinkage

Erik van Zwet[1] | Simon Schwab[2,3] | Stephen Senn[4]

[1]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

[2]Center for Reproducible Science (CRS), University of Zürich, Zürich, Switzerland

[3]Epidemiology, Biostatistics and Prevention Institute (EPBI), University of Zürich, Zürich, Switzerland

[4]Statistical Consultant, Edinburgh, UK

Correspondence
Erik van Zwet, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands.
Email: E.W.van_Zwet@lumc.nl

Abstract

We abstract the concept of a randomized controlled trial as a triple (β, b, s), where β is the primary efficacy parameter, b the estimate, and s the standard error $(s > 0)$. If the parameter β is either a difference of means, a log odds ratio or a log hazard ratio, then it is reasonable to assume that b is unbiased and normally distributed. This then allows us to estimate the joint distribution of the z-value $z = b/s$ and the signal-to-noise ratio $\text{SNR} = \beta/s$ from a sample of pairs (b_i, s_i). We have collected 23 551 such pairs from the Cochrane database. We note that there are many statistical quantities that depend on (β, b, s) only through the pair (z, SNR). We start by determining the estimated distribution of the achieved power. In particular, we estimate the median achieved power to be only 13%. We also consider the exaggeration ratio which is the factor by which the magnitude of β is overestimated. We find that if the estimate is just significant at the 5% level, we would expect it to overestimate the true effect by a factor of 1.7. This exaggeration is sometimes referred to as the winner's curse and it is undoubtedly to a considerable extent responsible for disappointing replication results. For this reason, we believe it is important to shrink the unbiased estimator, and we propose a method for doing so. We show that our shrinkage estimator successfully addresses the exaggeration. As an example, we re-analyze the ANDROMEDA-SHOCK trial.

KEYWORDS

achieved power, Cochrane review, exaggeration, randomized controlled trial, type M error

1 | INTRODUCTION

It is nearly three quarters of a century since what is generally regarded as the first modern randomized clinical trial, the UK Medical Research Council study of the effectiveness of streptomycin in tuberculosis.[1] Since then, tens of thousands of randomized controlled trials (RCT) have been conducted. The purpose of this article is to study this wealth of information, and to try to learn from it.

We have collected the results of more than 20 000 RCTs from the Cochrane Database of Systematic Reviews (CDSR), which is the leading journal and database for systematic reviews in health care. These data allow us to determine the

broad statistical properties of RCTs. In particular, we are able to estimate the distribution of the achieved power across all RCTs. We find that the achieved power is often quite low. Low statistical power has been noticed before in specific domains of biomedical research, see for instance.[2,3]

The fact that achieved power is typically low does not imply that the usual sample size calculations aiming for 80% or 90% power are wrong. The goal of such calculations is to guarantee high power against a particular alternative that is considered to be of clinical interest. The fact that high power is often not achieved is merely an indication that treatments often do not provide the benefit that was hoped for.

Low power has important implications for the interpretation of the results of a given trial. It is well known that conditional on statistical significance, the estimate of the effect size is positively biased. This bias is sometimes called the "winner's curse", and it is especially large when the power is low.[4-6] With the Cochrane data we can quantify the bias quite precisely.

To be more specific, we will represent an RCT by a triple (β, b, s) where β is the primary effect parameter and b is an unbiased, normally distributed estimator of β, with standard error s. The main idea of the article is that we can use the observed pairs (b, s) from the RCTs in the Cochrane database to estimate the *joint* distribution of the z-value $z = b/s$ and the signal-to-noise ratio $\mathrm{SNR} = \beta/s$. We can use this joint distribution to calculate the distribution of many interesting statistical quantities, conditional on the observed z-value. For example, we can compute the conditional coverage of the usual 95% confidence interval, given z. Or we can compute the conditional median of the exaggeration $|b/\beta|$, given z.

We realize that some readers may consider the very notion of a probability distribution for the β's to be inherently Bayesian. However, non-Bayesians may wish to think of this distribution as just an empirical model for the true effects across all trials, and in order to understand the variation of estimates in relation to putative parameter values invoking a (joint) probability model is a powerful tool.

Applying our method, we find considerable overestimation of the effect and under-coverage of the confidence interval when the z-value exceeds 1.5. These are serious issues which are undoubtedly a part of the explanation for the phenomenon of poor replication.[2,4,7,8] In Section 4 we note that the joint distribution of the SNR and the z-value allows us to compute the conditional expectation $\mathbb{E}(\mathrm{SNR} \mid z)$. Since $\beta = s \cdot \mathrm{SNR}$, this immediately suggest $s \cdot \mathbb{E}(\mathrm{SNR} \mid z)$ as a (shrinkage) estimator for β. We find that this new estimator almost entirely eliminates the problem of overestimation.

In summary, the purpose of the article is 2-fold. First, we analyze some key statistical properties of the population of RCTs in the Cochrane database, such as the achieved power, the exaggeration, and the coverage. Second, we propose a novel shrinkage estimator to address the exaggeration. We conclude this article with a short discussion.

2 | STATISTICAL ANALYSIS OF A COLLECTION OF RCTS

In this article, we abstract the concept of a RCT as a triple (β, b, s), where β is the primary efficacy parameter, b the estimate and s the standard error ($s > 0$). The parameter β is either a difference of means, a log odds ratio or a log hazard ratio. We will ignore small sample issues by assuming that b is a normally distributed, unbiased estimator of β with known standard error s. In other words, we will make the following assumption.

Assumption 1. Conditionally on β and s, b is normally distributed with mean β and SD s.

While this assumption may appear to be overly simplistic, we emphasize that inference based on Wald type confidence intervals and associated P-values is very common. When the sample size is not too small, say at least 60, this is quite appropriate.

Of course, we never observe the true effect β but Assumption 1 implies that we can use observed pairs (b_i, s_i) $(i = 1, 2, \ldots, n)$ to estimate the *joint* distribution of the z-value $z = b/s$ and the signal-to-noise ratio $\mathrm{SNR} = \beta/s$. This works as follows. We start by estimating the marginal distribution of the z-value directly from the observed sample $z_1 = b_1/s_1, \ldots, z_n = b_n/s_n$. By Assumption 1, the z-value is the sum of the SNR and independent standard normal "noise". This means that the distribution of the z-value is the convolution of the distribution of the SNR and the standard normal distribution. Therefore, we can obtain the marginal distribution of the SNR by deconvolution of the estimated density of the z-value and the standard normal density. Now we have the marginal distributions of both the z-value and the SNR. Assumption 1 tells us that the conditional distribution of the z-value given the SNR is normal with mean SNR and SD 1. Putting it all together, we have the joint distribution of the z-value and the SNR.

We start with the estimation of the marginal distribution of the z-value. We have obtained pairs (b_i, s_i) from 23 551 RCTs from the Cochrane Database of Systematic Reviews (CDSR). These derived data are publicly available at Reference 9. Each
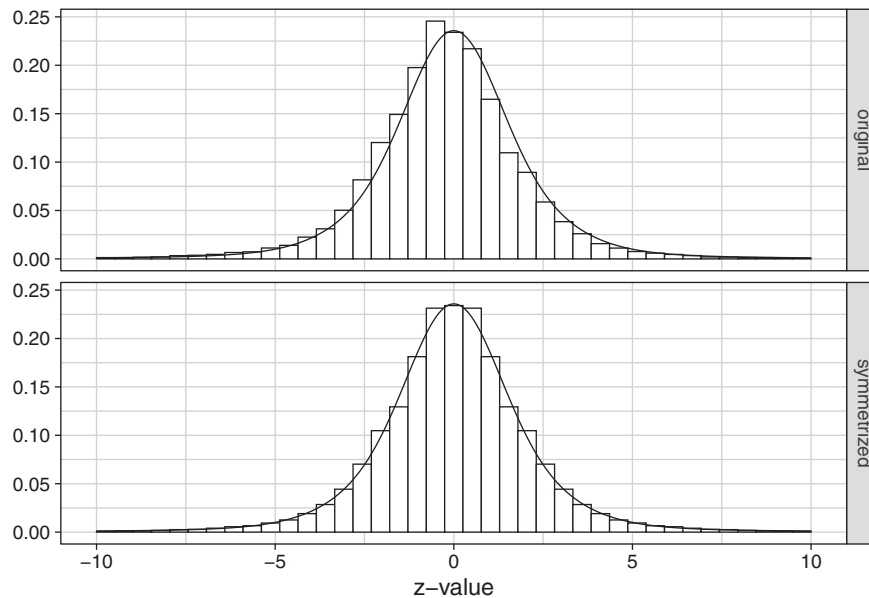
FIGURE 1 Top panel: The histogram of the observed z-values together with our fit based on a mixture of 4 zero-mean normal distributions. Bottom panel: The symmetrized histogram together with the same fit

TABLE 1 Estimated 4-part normal mixture distributions of the z-value and the SNR

	Comp. 1	Comp. 2	Comp. 3	Comp. 4
Proportions	0.32	0.31	0.30	0.07
SD of the z-value	1.17	1.74	2.38	5.73
SD of the SNR	0.61	1.42	2.16	5.64

pair belongs to a different study, and we have tried to obtain the estimate and standard error of the first or primary effect. The R code to generate the data and additional details about the data collection are provided in Reference 10. Briefly, for each systematic review the data were downloaded from the Cochrane Library as an XML file, parsed, and aggregated in a large database in R. The database included the sample sizes, mean and SD (for continuous outcomes) or the number of events (for dichotomous outcomes) for both the intervention and control arm. Intervention effects and standard errors were estimated using mean differences for continuous outcomes, and odds ratios for dichotomous outcomes. In order to identify which studies were RCTs, parsing of all the study characteristics tables on the Cochrane Library website was required and the database was extended with this information.

The histogram of the observed z-values is shown in the top panel Figure 1. We see that it is nearly symmetric, but not quite. This (near) symmetry was also observed by Djulbegovic et al who conducted a large meta-analysis of 860 published and unpublished phase III RCTs.[11] We prefer not to use the information about the sign of the z-values, for two reasons. First, using this information implies that we would treat the active treatment differently from control in our inferences. In most settings, that would not be acceptable. Second, many of the results that we present here depend on the z-value only through its absolute value.

Disregarding the sign, the *symmetrized* histogram of the z-values is shown in the bottom panel of Figure 1. We constructed it by using the R command `hist(c(-z,z))`. The symmetrized histogram of the z-values is symmetric and unimodal. Such distributions can be well approximated by a mixture of zero-mean Gaussians. We have used the R package "flexmix",[12] which implements the EM algorithm, to estimate it. We tried 1 up to 6 components, and found that more than 4 components made no discernible difference to the estimated distribution. The estimated mixture density fits the original histogram of the z-values reasonably well. The fit to the symmetrized histogram is excellent. The estimates of the 4 mixture components are given in Table 1.
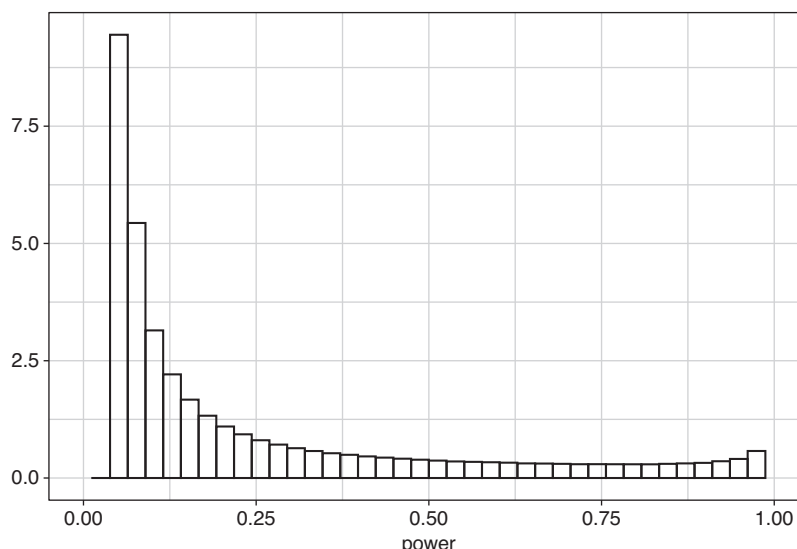
F I G U R E 2 Histogram of a sample of size 10^6 from the estimated distribution of the power

Recalling that the z-value is the sum of the SNR and independent standard normal noise, we can obtain the distribution of the SNR by deconvolution. Since the distribution of the z-value is a mixture of normal distributions, this is particularly easy; we simply subtract 1 from the variances of the components. The resulting SDs are given in Table 1.

By Assumption 1, the conditional distribution of the z-value given the SNR is normal with mean SNR and SD 1. The conditional distribution of the SNR given the z-value is slightly more complicated, but it can be obtained by straightforward application of Gaussian theory. It is again a mixture of Gaussians, and in the Appendix we provide the relevant formulas and a few lines of R code to compute it. We demonstrate the code in Section 4.3.

3 | POWER, EXAGGERATION, AND COVERAGE

Many statistical quantities depend on (β, b, s) only through the z-value and the SNR. Since we now have an estimate of their joint distribution across all the RCTs in the Cochrane database, we can quantify the achieved power, exaggeration (ie, overestimation of the magnitude of β), and coverage.

3.1 | Power

The achieved power of the two-sided Wald test of $H_0 : \beta = 0$ at level 5% depends on the SNR and is given by

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

where Φ is the cumulative distribution function of the standard normal distribution. The power is an even function, so it depends on the SNR only through its absolute value. In fact, the power is a strictly increasing function of absolute value of the SNR. Using formula (1), it is easy to transform a sample from the distribution of the SNR into a sample from the distribution of the power. We generated such a sample of size 10^6 and show the histogram in Figure 2. In Table 2 we report a number of quantiles of the distribution of the power. For example, the median of the power is 13%. The average power is 28%. O'Hagan et al[13] discuss the importance of average power to which they refer as the "assurance".

Most RCTs are designed to have 80% or 90% power against an alternative that is considered to be of clinical interest. As we can see from Figure 2 and Table 2, the *achieved* power is usually much lower than 80%. Indeed, we estimate that about 88% of RCTs have power less than 80%. However, this does not mean that most sample size calculations are mistaken. It merely indicates that many treatments do not have the effect that was considered to be of clinical interest. Finding good treatments is not easy! Other factors may also contribute to low power, such as difficulties with subject recruitment or

TABLE 2 Estimated quantiles of the absolute value of the signal-to-noise ratio and the achieved power

	Q10	Q25	Q50	Q75	Q90
|SNR|	0.14	0.36	0.83	1.73	3.03
Power	0.05	0.07	0.13	0.41	0.86
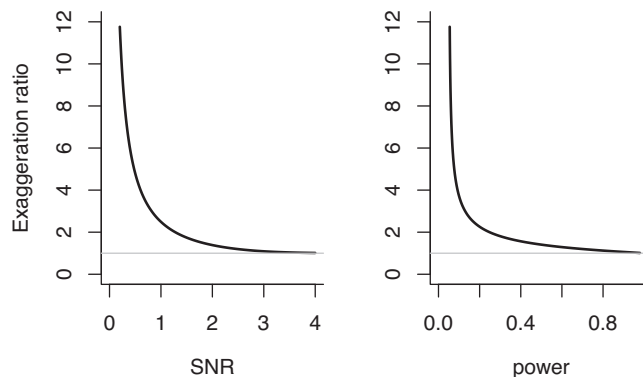
FIGURE 3 These curves show the mathematical relation between the exaggeration ratio (expected type M error) as a function of the SNR and the power

larger between-subject variation than anticipated. In any case, the fact remains that the achieved power in the majority RCTs is quite low, and this has consequences for our inferences.

3.2 | Exaggeration

Another important statistical quantity is the ratio $R = |b|/|\beta|$. Its expectation $\mathbb{E}(R \mid \beta, s)$ could be called the relative bias of the magnitude. In the absence of bias, it is equal to one. However, the fact that b is unbiased for β implies that $|b|$ is positively biased for $|\beta|$ (Jensen's inequality). So the unbiasedness of b implies that $\mathbb{E}(R \mid \beta, s)$ is always greater than one.

Gelman and Carlin[5] define the "expected type M error" or "exaggeration ratio" as

$$\mathbb{E}(R \mid \beta, s, |b|/s > 1.96). \tag{2}$$

It represents the factor by which the magnitude of the effect may be expected to be overestimated when we condition on significance at the 5% level (two-sided). This exaggeration is sometimes referred to as the winner's curse. Undoubtedly, the winner's curse is to a considerable extent responsible for disappointing replication results.

It turns out that the exaggeration ratio depends on β and s only through the absolute value of the SNR. In fact, van Zwet and Cator[6] show that the exaggeration ratio is decreasing in |SNR|. Since the power is a strictly increasing function of |SNR|, it follows the exaggeration ratio is also a decreasing function of the power, see also References 4,5. In Figure 3 we show the exaggeration ratio as a function of |SNR| and as a function of the achieved power. If the achieved power is 13% (which is the median across the RCTs in the Cochrane database), then the exaggeration is about 3. This means that if a study with median power reaches significance, we would expect the magnitude of the effect to be overestimated by a factor of 3.

Of course, in practice we do not know β so we know neither the achieved power nor the exaggeration ratio. However, note that R is just a function of the SNR and the z-value

$$R = \frac{|b|}{|\beta|} = \frac{|b|/s}{|\beta|/s} = \frac{|z|}{|\mathrm{SNR}|}. \tag{3}$$

Since we have the joint distribution of the SNR and the z-value, we can calculate the conditional distribution of R given z. We represent the conditional distribution of R given z by its three quartiles in the left panel of Figure 4.
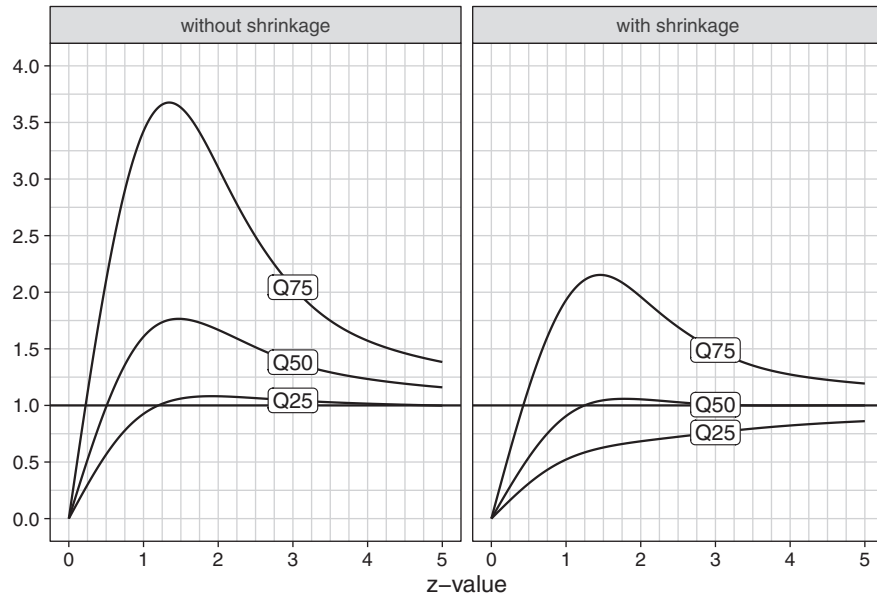
FIGURE 4 Left panel: The distribution of the ratio $R = |b|/|\beta|$ conditional on the z-value, represented by its three quartiles. Right panel: The distribution of the ratio $|\widehat{\beta}|/|\beta|$ conditional on the z-value

We notice that considerable bias is already present at fairly small values of z. When $z = 1.96$, the conditional median of R is about 1.7.

3.3 | Coverage

We conclude this section by considering the conditional coverage of the usual 95% confidence interval given the observed z-value. Note that

$$P(b - 1.96\ s < \beta < b + 1.96\ s \,|\, z) = P(z - 1.96 < \text{SNR} < z + 1.96 \,|\, z). \tag{4}$$

Again, the joint distribution of the SNR and the z-value allows us to calculate this expression. We plot the conditional coverage in Figure 5. We see that if z exceeds 2, the conditional coverage is much lower than the nominal level of 95%.

4 | SHRINKAGE

When the z-value exceeds about 1.5, we see substantial overestimation of the magnitude of the effect in Figure 4 and undercoverage of the confidence interval in Figure 5. These are clearly very serious problems. In this section we propose a novel, data-based shrinkage estimator by conditioning on the observed z-value.

4.1 | A shrinkage estimator

We have estimated the conditional distribution of the SNR given the z-value, so we also have the conditional expectation $\mathbb{E}(\text{SNR} \,|\, z)$. Since $\beta = s \cdot \text{SNR}$, this immediately suggests an alternative estimator for β, namely

$$\widehat{\beta} = s \cdot \mathbb{E}(\text{SNR} \,|\, z). \tag{5}$$

Since the marginal distribution of the SNR is a mixture of zero-mean normal distributions, $\widehat{\beta}$ will be "pulled" towards zero so that $|\widehat{\beta}|$ will be less than $|b|$. In other words, $\widehat{\beta}$ is a *shrinkage* estimator. We stress that the estimator does not
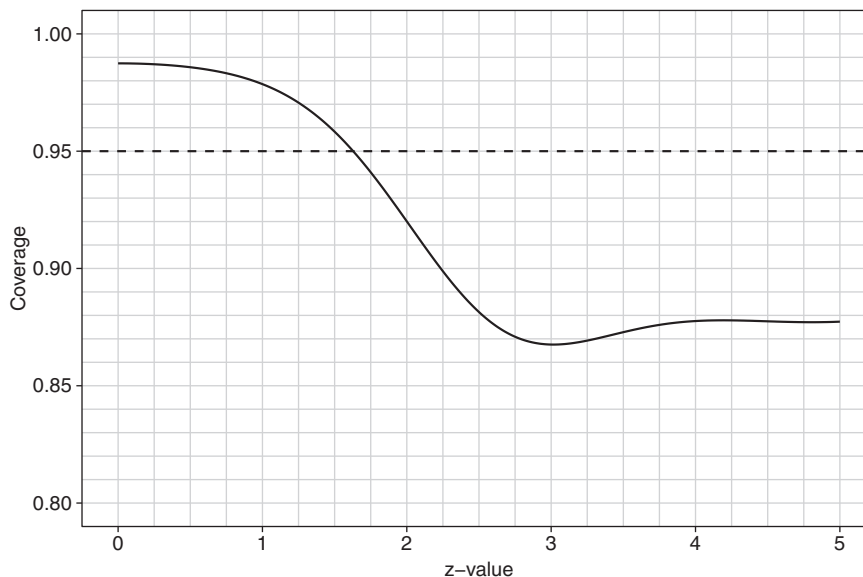
FIGURE 5   Probability that the interval $[b - 1.96\,s, b + 1.96\,s]$ covers β conditional on the z-value

apply a universal shrinkage factor. More shrinkage will be applied to noisy estimates with small z-values, while almost no shrinkage is applied to estimates with large z-values. In this sense, the amount of shrinkage is adaptive to the amount of information in the data.

We can see in the left panel of Figure 4 that $|b|$ is typically very biased for $|\beta|$. In the right panel we show the three quartiles of the conditional distribution of $|\widehat{\beta}|/|\beta|$ given z. We see that the shrinkage successfully addresses the exaggeration.

4.2 | Independence assumption

Until now, we carried out all our inferences conditionally on the z-value. However, we observe both b and s, so we are ignoring some potentially relevant information. In particular, we may *not* conclude that our shrinkage estimator $\widehat{\beta}$ is equal to the Bayes estimator (posterior mean) $\mathbb{E}(\beta \mid b, s)$.

Assumption 1 in Section 2 actually makes it possible to estimate the complete joint distribution of (β, b, s) from a sample of pairs $(b_1, s_1), \ldots, (b_n, s_n)$. However, modeling the three-dimensional distribution of (β, b, s) is much harder than the two-dimensional distribution of (z, SNR), especially since z and SNR have such a simple relation to each other. The following Proposition formalizes when the joint distribution of z and SNR is sufficient for computing the conditional distribution of β given (b, s).

Proposition 1. *Let $f_{\beta|b,s}$ denote the conditional density of β given (b, s) and $f_{\mathrm{SNR}|z}$ denote the conditional density of SNR given z. If the pair (b, s) and the SNR are conditionally independent, given the z-value, then*

$$f_{\beta|b,s}(x) = \frac{1}{s} f_{\mathrm{SNR}|z}(x/s). \tag{6}$$

Proof. Let $f_{\mathrm{SNR}|b,s}$ denote the conditional density of SNR given (b, s). Since $\beta = s \cdot \mathrm{SNR}$, we have

$$f_{\beta|b,s}(x) = \frac{1}{s} f_{\mathrm{SNR}|b,s}(x/s). \tag{7}$$

The conditional independence condition now says

$$f_{\mathrm{SNR}|b,s}(x) = f_{\mathrm{SNR}|z}(x), \tag{8}$$

and the result follows.   ∎

Under the condition of Proposition 1, the conditional distribution of β given (b, s) is equal to the conditional distribution of the SNR given z, scaled by a factor s. In particular,

$$\mathbb{E}(\beta \mid b, s) = s \cdot \mathbb{E}(\text{SNR} \mid z) = \widehat{\beta}. \quad (9)$$

Moreover, from the conditional distribution of the SNR given z, we can construct an interval such that the conditional probability of covering the SNR is 95%. Under the condition of Proposition 1 we can simply scale this interval by s to get an interval for β which has 95% coverage, conditionally on z.

The distribution of the z-value across the Cochrane database carries only information about the distribution of the SNR, which in turn carries only information about the distribution of the achieved power. So, if we use the observed z-value as a basis for inference about β via (6), then we are using only information about the distribution of the achieved power across the Cochrane database. We are ignoring any additional information that may be present in the pair (b, s).

Doing inference about β via (6) has one more practical advantage. Scaling by s means that the inference about β is equivariant under changes of the unit of measurement of the outcome. In other words, it does not matter for our conclusions if the outcome happens to be measured in grams or kilograms, for example.

4.3 | Example

To demonstrate our shrinkage estimator, we will consider the recent ANDROMEDA-SHOCK RCT.[14] The aim of this trial was to determine if treating patients on the basis of so-called capillary refill time (CRT) is superior to treating them on the basis of blood lactate levels. The primary outcome was survival at 28 days and the primary effect parameter was the hazard ratio which was estimated at 0.75 with 95% confidence interval from 0.55 to 1.02. This is just shy of statistical significance at the 5% level, which led the investigators to conclude that the experimental strategy "did not reduce all-cause 28-day mortality".

We can relate the results of this particular RCT to what we know about RCTs in general. This means that we must view the ANDROMEDA-SHOCK trial as a "typical" RCT in the sense that its signal-to-noise ratio is exchangeable with that of the other RCTs in the Cochrane database. In other words, we use *only* the information that ANDROMEDA-SHOCK trial is an RCT. We use none of the particular features of the trial such as the disease, the treatment, the population, the sample size, the type of outcome, et cetera. Moreover, we also treat the two arms of the trial exchangeably.

The estimated log hazard ratio is $b = \log(0.75) = -0.29$ with 95% confidence interval from -0.6 to 0.02. From this confidence interval, we can compute the standard error of b to be $s = 0.16$. The z-value is $-0.29/0.16 = -1.81$. We can now run the code below to compute our shrinkage estimator of the log hazard ratio. Note that the function `conditional()`, which is provided in the Appendix, must be sourced first.

```
s <- 0.16
p <- c(0.32,0.31,0.3,0.07)              # from Table 1
tau <- c(0.61,1.42,2.16,5.64)           # from Table 1
res <- conditional(z=-1.81,p=p,tau=tau) # conditional distr. of the SNR given z
                                          (Appendix)
s * sum(res$q * res$m)                   # betahat = s*E(SNR | z)
```

We find $\widehat{\beta} = -0.18$ and notice the considerable shrinkage compared to $b = -0.29$. Next, we can also construct an interval that covers the SNR with probability 95% conditionally on the observed z-value. Computing the quantiles of a one-dimensional normal mixture distribution is not difficult. It can be done via a root-finding algorithm, but we can also use the function `qmixnorm` from the R package "KScorrect".

```
KScorrect::qmixnorm(p=c(0.025,0.975),mean=res$m,sd=res$sigma,pro=res$q,expand=0.5)
```

We find the 95% interval from −3.0 to 0.43 for the SNR. We multiply this interval by s to get an interval for β which also has 95% coverage conditionally on z. In this way, we find the interval from −0.48 to 0.07 for β . Finally, we can also compute the conditional probability that β is negative, given the observed z-value.

```
sum(res$q * pnorm(0,mean=res$m,sd=res$sigma))
```

We find that this probability is 91%. This is in line with a Bayesian re-analysis of the ANDROMEDA-SHOCK trial.[15]

5 | DISCUSSION

We have analyzed the primary results of a collection of more than 20 000 RCTs from the Cochrane Database of Systematic Reviews (CDSR). By estimating the joint distribution of the SNR and the z-value, we have been able to study the distribution of the achieved power. We have found that the achieved power is often very low. We have demonstrated how this leads to overestimation and undercoverage when the observed z-value exceeds about 1.5. We propose a shrinkage estimator that addresses these issues by conditioning on the observed z-value. The estimator is easy to compute from the reported effect estimate and its standard error. We have demonstrated its use by re-analyzing the ANDROMEDA-SHOCK trial.

Our shrinkage estimator allows us to take into account the well-known fact that many RCTs have low achieved power. We are not suggesting that our estimator should replace other approaches, but we do believe that it provides a useful addition to whichever pure Bayesian or frequentist alternative the reader might prefer as their first choice of analysis based on their own philosophical preference. We expect that Bayesians would want to use more prior information than the distribution of the achieved power across all RCTs. See Reference 16 for an extensive and balanced discussion of Bayesian approaches to RCTs. On the other hand, frequentists would prefer not to use any prior information at all.

We are not the first to use the Cochrane database to obtain general information about RCTs. Turner et al[17] and Rhodes et al[18] used the database to get information about the heterogeneity of effects within a meta-analysis of a particular treatment. We do believe that we are the first to propose using the database to obtain prior information about the SNR, or equivalently, the achieved power across all RCTs.

While the Cochrane collaboration puts much effort in collecting all available studies on a particular topic (both published and unpublished), it is likely that the database suffers from at least some publication bias, file drawer effects, fishing, forking paths etc. Unfortunately, this means that the true (achieved) power is likely to be even lower and the exaggeration even greater.

Strictly speaking, our analysis concerns only the Cochrane database. However, we believe that problems of low power and exaggeration are similar—if not worse—in many other areas of research.

ORCID

Erik van Zwet https://orcid.org/0000-0001-5537-3179
Simon Schwab https://orcid.org/0000-0002-1588-2689
Stephen Senn https://orcid.org/0000-0002-7558-8473

REFERENCES

1. Tuberculosis Trials Committee iMRCS. Streptomycin treatment for pulmonary tuberculosis. *BMJ*. 1948;2(4582):769-782.
2. Button KS, Ioannidis JP, Mokrysz C, et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci*. 2013;14(5):365.
3. Dumas-Mallet E, Button KS, Boraud T, Gonon F, Munafò MR. Low statistical power in biomedical science: a review of three human research domains. *Royal Soc Open Sci*. 2017;4(2):160254.
4. Ioannidis JP. Why most discovered true associations are inflated. *Epidemiology*. 2008;19(5):640-648.
5. Gelman A, Carlin J. Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. *Perspect Psychol Sci*. 2014;9(6):641-651.

6. van Zwet EW, Cator EA. The winner's curse and the need to Shrink. *Stat Neerl.* 2021;1-15.

7. Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. *JAMA.* 2005;294(2):218-228.

8. Button KS, Ioannidis JP, Mokrysz C, et al. Empirical evidence for low reproducibility indicates low pre-study odds. *Nat Rev Neurosci.* 2013;14(12):877-877.

9. Schwab S. Re-estimating 400,000 treatment effects from intervention studies in the Cochrane database of systematic reviews [Data set]. *Open Sci Framework.* 2020. https://doi.org/10.17605/OSF.IO/XJV9G

10. Schwab S, Kreiliger G, Held L. Assessing treatment effects and publication bias in the Cochrane database of systematic reviews across different specialties in medicine; 2021. https://osf.io/preprints/metaarxiv/t6cjx/.

11. Djulbegovic B, Kumar A, Glasziou P, Miladinovic B, Chalmers I. Trial unpredictability yields predictable therapy gains. *Nature.* 2013;500(7463):395-396.

12. Leisch F. FlexMix: a general framework for finite mixture models and latent class regression in R. *J Stat Softw.* 2004;11(8):1-18. http://www.jstatsoft.org/v11/i08/

13. O'Hagan A, Stevens JW, Campbell MJ. Assurance in clinical trial design. *Pharm Stat J Appl Stat Pharm Ind.* 2005;4(3):187-201.

14. Hernández G, Ospina-Tascón GA, Damiani LP, et al. 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. *JAMA.* 2019;321(7):654-664.

15. Zampieri FG, Damiani LP, Bakker J, et al. Effects of a resuscitation strategy targeting peripheral perfusion status versus serum lactate levels among patients with septic shock. a Bayesian reanalysis of the ANDROMEDA-SHOCK Trial. *Am J Respir Crit Care Med.* 2020;201(4):423-429.

16. Spiegelhalter DJ, Freedman LS, Parmar MK. Bayesian approaches to randomized trials. *J R Stat Soc A Stat Soc.* 1994;157(3):357-387.

17. Turner RM, Davey J, Clarke MJ, Thompson SG, Higgins JP. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane database of systematic reviews. *Int J Epidemiol.* 2012;41(3):818-827.

18. Rhodes KM, Turner RM, White IR, Jackson D, Spiegelhalter DJ, Higgins JP. Implementing informative priors for heterogeneity in meta-analysis using meta-regression and pseudo data. *Stat Med.* 2016;35(29):5495-5511.

APPENDIX

Here we discuss the conditional distribution of the SNR given the z-value. Now, suppose the SNR is distributed as a mixture of k zero-mean normal components with mixture proportions $p_1, \ldots, p_k$ and SDs $\tau_1, \ldots, \tau_k$. Then z is also a mixture of k zero-mean normals with the same mixture proportions, but SDs $\sqrt{\tau_i^2 + 1}$. The density of z is

$$g(z) = \sum_{i=1}^{k} \frac{p_i}{\sqrt{\tau_i^2 + 1}} \, \varphi \left(z / \sqrt{\tau_i^2 + 1} \right), \tag{A1}$$

where φ is the standard normal density function.

The conditional distribution of the SNR given z is again a mixture of normals. The mixing proportions are

$$q_i = \frac{p_i}{\sqrt{\tau_i^2 + 1}} \varphi \left(z / \sqrt{\tau_i^2 + 1} \right) \frac{1}{g(z)}, \tag{A2}$$

the means of the components are

$$m_i = \frac{z \, \tau_i^2}{\tau_i^2 + 1}, \tag{A3}$$

and the variances of the components are

$$v_i^2 = \frac{\tau_i^2}{\tau_i^2 + 1}. \tag{A4}$$

Here we present an R function to compute the conditional distribution of the SNR given the z-value when the marginal distribution of the SNR is a mixture of zero-mean Gaussians with a vector of mixing proportions p and associated SDs tau. The conditional distribution is again a Gaussian mixture with proportions q, means m and SDs sigma. The usage of the function is demonstrated in Section 4.3.

```
conditional <- function(z,p,tau) {
 tau2 <- tau^2
 q <- p*dnorm(z,0,sqrt(tau2+1))
 q <- q/sum(q)              # conditional mixing probs
 m <- z*tau2/(tau2+1)       # conditional means
 v <- tau2/(tau2+1)         # conditional variances
 sigma <- sqrt(v)           # conditional std devs
 data.frame(q,m,sigma)
}
```