THE MAIMED MARTIAN, CREDIBLE INTERVALS AND BIAS AGAINST BENEFIT

by David Kault

College of Science and Engineering, James Cook University, Townsville, Queensland, 4816, Australia Adjunct snr lecturer in mathematics and General Practitioner

Email: David.Kault@jcu.edu.au

Key Words:
Credible Intervals
Bayesian statistics
False negative results
Publication bias

Preprint of paper accepted for publication in the journal Evidence-Based Medicine (ebm.bmj.com) - one of the BMJ journals. Anticipated date of publication January 2017

ABSTRACT

A sad little story about a maimed Martian astronaut is used to illustrate a method of improving confidence interval calculations. Confidence intervals in medical statistics are currently calculated from the data available in a clinical trial or meta-analysis considered in isolation from all other information available on earth. Likewise, the Martian in the story uses only information available to her, in isolation from further information from earth. However there is further objective knowledge available to people on earth to improve the Martian's estimate. In the same way we have objective prior knowledge available to us outside of the current clinical trial results which we can use to improve confidence interval calculations. This prior knowledge is incorporated into the confidence interval calculations using Bayesian methods. The objective prior knowledge that is available is the fact that there were researchers who felt it worthwhile to conduct the trial and journal editors who felt it worthwhile publishing the results. It is shown here that the use of this information contracts the width of the log confidence interval by a factor of about 3/4 on average. Unlike standard confidence intervals, these new intervals also have the advantage of being directly interpretable in terms of probabilities. These probabilities also enable calculation of improved point estimates.

These calculations are applied to 100 randomly selected Cochrane systematic reviews, and show serious problems in assessing medical treatments. For treatments not involving new drugs or devices, it is shown that there is evidence of a bias towards a negative assessment. The calculations here make a quantitative adjustment for publication bias. They show that the proportion of negative assessments do not reflect an appropriate adjustment for publication bias.

INTRODUCTION

This paper extends a recent approach [1] to the analysis of clinical trial statistics, to give credible intervals. These improve the assessment of a specific clinical trial statistic by factoring in experience with clinical trial statistics from across the Cochrane Collaboration [2]. These credible intervals have a similar role to confidence intervals, but have several advantages – they are usually narrower, unlike confidence intervals they can be directly interpreted as giving probabilities and they automatically allow for publication bias. Software to calculate them is being made freely available. Credible intervals are calculated for 100 randomly selected studies from the Cochrane reviews. The results suggest that Cochrane tends to have a bias against finding a reasonable possibility of benefit.

In this paper, reviews of new drugs and new medical devices are excluded and the paper focusses on other treatments. This exclusion is motivated by concern

that publication bias may be unpredictably large where there are commercial pressures [3-6] and recognition that a conservative approach by Cochrane here, may be appropriate to allow for unexpected dangerous side effects.

To explain the idea behind the method, consider a quirky but sad tale of the first astronaut to come from Mars. The astronaut has the misfortune to arrive on earth in the dark and lands next to a man who is a violent anti-alien extremist. She is blinded in an attack by the extremist but is then rescued and taken to the police to give a statement about her attacker. The astronaut is aware of the metric system but she has no conception about the height of a typical human. She tells the police that given it was dark and that she got just a glimpse of her attacker, she is unsure of his height, she thinks he was around 200cm and is 95% confident that he was somewhere in the range 170cm to 230cm. The police however know the distribution of human male heights in their country and know that 95% are between 150cm and 190cm and virtually no-one is anywhere near 230cm. It is therefore reasonable for them to believe that it is far more likely that the attacker has a height somewhere around the lower half of the 170 to 230 range. In fact the police do better and as explained in Appendix 1, they do a calculation using Bayes' theorem, to find that with the information available, there is a 95% chance that the attacker's height is in the range 163cm to 196cm. The calculation results in the width of the interval giving the likely height of the attacker being contracted from 60cm to 33cm. Equations (1) and (2) of Appendix 1 show that regardless of the accuracy of the Martian's estimate and regardless of the variability of human height, there will always be at least some contraction of the width of the confidence interval. It will also be shifted towards a plausible range according to a weighted average of the Martian's guess and the average human height with the weights being determined by the uncertainty in the Martian's guess and the variability in human height.

The situation with clinical trials and meta-analyses is similar in a sense to the situation of the Martian. If we stop our analysis after calculating a confidence interval we are using no more information about our treatment than the Martian used about her attacker's height. However, we can improve the situation by using the sort of calculations done by the police who rescued the Martian and factor in objective background knowledge. Objective background knowledge is available - if we look at all the clinical trials and meta-analyses that have ever been done and recorded in the Cochrane Collaboration [2], there is a particular pattern or distribution for effectiveness, just as there is a particular distribution for human heights. In effect, we use the objective background knowledge that we are looking at the results of a clinical trial of a treatment where there was sufficient reason to trial that treatment in the first place.

Improving uncertain effectiveness estimates using the distribution of effectiveness

is a bit more complicated than improving uncertain height estimates using the distribution of heights, for several reasons:

- The distribution of human heights is available from a large number of quite accurate measurements of human height whereas the distribution of effectiveness is available only from a limited sample of uncertain measurements [1].
- The distribution of human heights follows a normal distribution quite closely, whereas the distribution of effectiveness is not as straightforward.
- In compiling the distribution of human heights there was no tendency to systematically exclude dwarves from the sample, whereas it can be assumed that there was a degree of publication bias operating to exclude "non-significant" measures of effectiveness from publication.

However, allowance can be made for these complications and an effectivenss distribution can be obtained. Effectiveness can be measured in several different ways. Some measures related to effectiveness are dependent on the unit of measurement (eg. kilograms weight loss versus pounds weight loss) or may reflect the size of the study as well as the effectiveness (eg. the p-value). A measure of effectiveness free of these problems, is readily available when dealing with dichotomous outcomes such as live or die and when these outcomes are summarised by statistics such as relative risks, relative hazards or odds ratios. For convenience we abuse language and use the term "relative risk" or RR in what follows to refer to all such statistics. With such statistics, a relative risk of 1.00 signifies zero effectiveness. Since log(1.00) is zero, this suggests a choice of log relative risk as the effectiveness measure. This choice of effectiveness measure is also convenient because conventional confidence interval calculation of relative risk assumes a normal distribution after a log transformation [7].

The relative risks used here have to satisfy certain criteria [1]. Relative risks are eligible if they come from trials or meta-analyses in which there is a comparison of a standard treatment with the same standard treatment plus an additional treatment. The rationale is that such a trial would not take place unless there was a general sentiment that the additional treatment might be of benefit and is unlikely to be harmful, whereas this would not apply in the less common situation where two mutually exclusive treatments are compared – such as two competing surgical approaches. Presumably, for eligible trials, the general sentiment will be partly reflected in reality. Treatments will often have at least some modest effectiveness, a few might be remarkably effective but some will be entirely useless with zero effectiveness and perhaps a few will be counterproductive. These ideas were embodied in a mathematical model which allocated a proportion to have precisely zero effectiveness with the remainder having an effectiveness chosen from a normal

distribution centred on a positive number. Data in the form of 101 effectiveness measures previously obtained from Cochrane [1] were used to calculate the three parameters of the model. One parameter gives the proportion of treatments that have zero effectiveness and the other two describe the normal distribution of effectiveness for the remaining treatments. The parameter estimates were made by curve fitting as explained elsewhere [1]. Allowance was then made for the uncertainty about the three estimated parameters by a bootstrapping method described in the appendix. Importantly, the method also makes allowance for a publication bias against non-significant results by a factor of 2.78 [8]. The outcome is a probability distribution of effectiveness of diverse treatments across all medical problems as shown in Figure 1.

This distribution can then be used in a similar way to the police use of the distribution of diverse heights of all men to improve the Martian's estimate. The mathematics here is given in the Appendix. The idea behind the mathematics can be loosely explained in words using as an example, a trial which yields a relative risk of 2.83 with a 95% confidence interval of (2.00, 4.00). We might ask, "What is the probability that the true relative risk is around 2.00?". From the definition of a 95% confidence interval, we know that if the true value was either 2.00 or 4.00 we would, in each case, have a probability of 2.5% of getting a relative risk of around 2.83 (actually 2.83 or beyond). Referring to Figure 1 shows that across Cochrane we estimate relative risks around 2.00 to be about 20 times ($\approx 0.406/0.0206$) more common than values around 4.00 . Combining this with the fact that in this clinical trial, values of both 2.00 and 4.00 are equally capable of giving the observed value of 2.83, tells us that a relative risk for this treatment of around 2.00 is about 20 times more likely than a value around 4.00. Taking into account the log scale and using the mathematics in the appendices and elsewhere [1, 9] makes these ideas more precise and extends them a little to give the probability of any value for the true effectiveness. This allows calculation of a credible interval that has a 95% chance of containing the value of the true amount of effectiveness. Free software is made available to implement these calculations. (It is intended that this will become available on Research Data Australia https://researchdata.ands.org.au/) In the example here, the calculations show that there is a 95% probability that the true relative risk in this trial is in the range (1.48, 3.41). The expected relative risk (E(RR)) is also calculated as described in the appendix. This is a probability weighted average of the likely effectiveness of a treatment. The expected relative risk is a modification of the relative risk which takes account of the background distribution of effectiveness and makes appropriate adjustment to compensate for the log transformation. In this example the E(RR) is 2.35. The calculations also tell us that, to two decimal places, the probability that the treatment is either counterproductive or ineffective is, in both cases, zero.

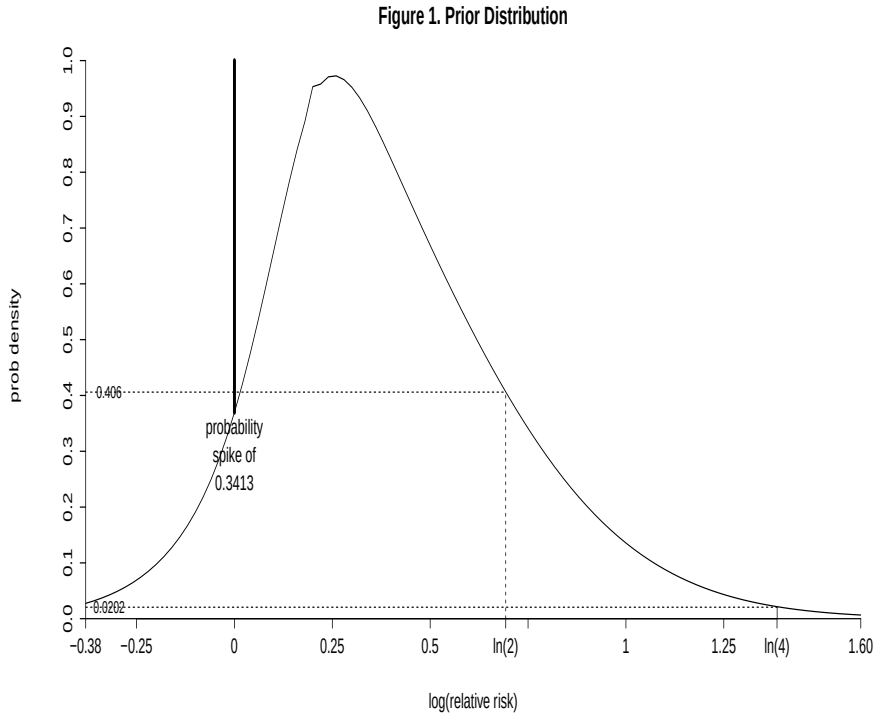

Figure 1: This gives the distribution of effectiveness as measured by the log of the relative risk. It shows the function given as equation (5) in the appendix. Dotted lines on the graph show the coordinates of two points:- the height of the graph at a log relative risk of ln(4) (the natural log of 4) is 0.0202 and the height of the graph at a log relative risk of ln(2) is about 20 times larger at 0.406. This tells us that across Cochrane, log relative risks around ln(2), are about 20 times more common than log relative risks around ln(4). On a scale of relative risks themselves, this translates into relative risks around 2 being about 40 times more common than relative risks around 4.

METHODS

A further 100 relative risks with their confidence intervals from trials and meta-analyses throughout Cochrane were selected using a random number generator and the protocol described previously [1]. As well, for each study selected, the review abstract was read and a ranking was given regarding the treatment. A rank of 1 or 2 was assigned if the reviewers thought that the treatment was definitely or probably of zero effectiveness (or counterproductive). A ranking of 3 was assigned if the review abstract indicated that there remained complete uncertainty about the effectiveness of the treatment and a rank of 4 or 5 was assigned if the review indicated that the treatment was probably or definitely effective.

For the effectiveness distribution to be defined, it is necessary that an anticipated effect is consistently reflected by the numerical measure of effectiveness. However, a relative risk bigger than 1 may be associated with a beneficial effect as for example in terms of a higher "risk" of giving up smoking with an anti-smoking treatment, or with a harmful effect as in a higher risk of dying with the additional treatment. For consistency, the algorithm described here must be given information about whether the direction of effect tends in an anticipated or opposite direction. Internally the algorithm stores all relative risks in favour of the extra treatment as a relative risk bigger than 1. In effect a relative risk of dying with the additional treatment of 0.5, is translated into a relative risk of 2.0 of dying if given just the standard treatment. The figures display results using this approach. However, for consistency with the original confidence intervals, Table 2 displays results as though there had been no inversion.

Occasionally, there is an additional twist here. Sometimes researchers are hoping for a negative answer to the question "Is standard treatment plus the additional treatment better than standard treatment alone?". For example, to reduce costs, researchers may hope to show that less highly trained staff can perform a procedure as well as more highly trained staff. Here one would regard "standard treatment plus the additional treatment" as the procedure being performed by someone with baseline training plus extra training and "standard treatment" as the procedure being performed by someone with baseline training alone. The researchers may be hoping to find "no difference" between the two arms of the trial, but the outcome that could be anticipated is that, if there is any effect, it will be that the extra training makes a beneficial difference - something that would be a disappointing result for the those conducting the trial.

RESULTS

A total of 340 Cochrane reviews were searched to give 100 relative risks. None were significantly counterproductive, 16 showed a non-significant negative effect, in 3 cases the effect size was precisely zero, 29 showed a non-significantly positive effect and 52 showed a significantly positive effect. This outcome is consistent ($\chi^2_8 = 9.31$, p-value = 0.32) with the outcome found when 100 and then 101 relative risks were collected from Cochrane in two previous studies [1, 9]. The 100 reviews with the relative risk used, are numbered and are listed in Table 1.

The effectiveness assessment rankings 1, 2, 3, 4, 5 (that is, from certainly ineffective or counterproductive to certainly effective) were appropriate to 6, 7, 28, 31 and 28 Cochrane reviews respectively. It is of interest that of the reviews with the rank of 1 there was no negative RR result and 5 positive RR results, for rank 2 the figures were 2 and 4, for rank 3 the figures were 11 and 17, for rank 4 the figures were 2 and 28 and for the most positive effectiveness ranking of 5 the figures were 1 and 27. All studies except for one ranked 3 (uncertain) had a positive E(RR), that is, an expected effect in the direction anticipated. The results are displayed in Figure 2.

The remaining information from the analysis is displayed in Table 2. For each study numbered in Table 1, Table 2 gives the original confidence interval for RR, together with the probabilities of counterproductive, zero or positive effect, the 95% credible interval, the E(RR) and the effectiveness ranking. In all cases with a low ranking of 1 or 2 (definitely or probably useless or counterproductive), the probability of a positive effect is more than 27% and in all cases exceeds the probability of a counterproductive effect by at least 18%.

On average, the width of the 95% credible intervals is about $\frac{3}{4}$ that of the original 95% confidence interval on the log scale (95% confidence interval for this value of 0.75 is (0.70, 0.81)).

DISCUSSION

It is to be expected that there will be an imperfect correlation between the RR or the E(RR) and the Cochrane assessment of effectiveness as summarised by the ranking used here. Unlike the E(RR) which intrinsically takes account of publication bias, Cochrane systematic reviews only occasionally take formal account of publication bias by presenting funnel plots. However, the E(RR) calculation relies only on the first statistic cited in a Cochrane review, whereas the review itself may take account of a number of analyses which may be evaluating a variety of different endpoints and the review will also judge the quality of the clinical trials. Cochrane reviews also occasionally take some account of observational

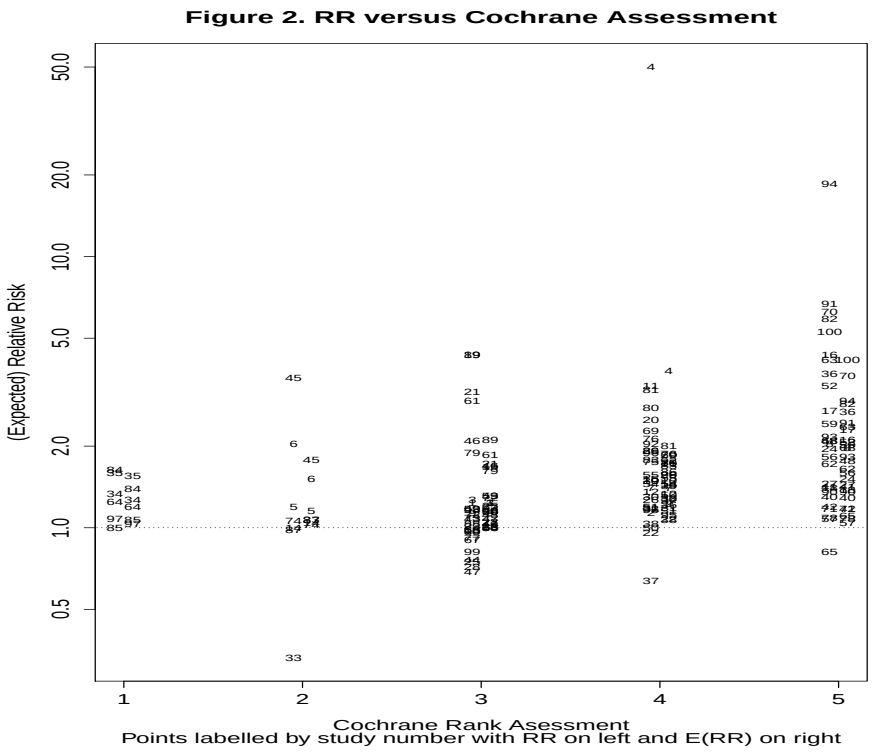

Figure 2: This gives some of the information of Table 2 in diagrammatic form. The 100 randomly chosen studies are grouped according to the Cochrane reviewers' assessment of their likely effectiveness. Above each effectiveness ranking from 1 to 5, the first relative risk mentioned in each of the 100 studies is plotted. The simple relative risk is displayed slightly to the left and the expected relative risk is displayed slightly to the right. The points are marked by small font numbers which refer to the study numbers given in Tables 1 and 2. Due to display limitations perhaps only the more unusual study numbers will be entirely legible.

studies and non-randomised trials and may make a somewhat subjective judgement about whether an outcome is clinically significant as well as statistically significant.

Nevertheless there is a disturbing discordance in the results between the effectiveness ranking on the one hand and the RR and even moreso, the E(RR) on the other. In particular, the results show that in terms of conventional RR's, E(RR)'s and probabilities, all 13 treatments ranked as definitely useless or counterproductive, deserved a higher ranking. It could be reasonable for many treatments with weakly positive RRs or E(RR)s to be ranked as inconclusive, perhaps because they were underpowered relative to the effect size or the trial had design faults, but a conclusion stating that the results show the treatment effects are probably or convincingly negative, seems wrong. It is possible to explain this discordance by arguing that my rankings are inappropriate, or the effect size is convincingly too small to be worthwhile, or that bias in a clinical trial was so large that one would anticipate a negative result from an unbiased study, or that second or subsequent statistics in a review gave a convincingly negative impression. However, perusal of the negatively ranked reviews do not support such arguments (except perhaps for review 74). Instead, it seems almost certain that the negative reviews in the context of mostly positive RRs represent an undue conservatism by the reviewers. Presumably, this conservatism reflects concern about publication bias. A significant finding of this paper is that this concern is misguided. As explained above, the E(RR)s make appropriate allowance for publication bias, and all the E(RR)s of the negatively ranked reviews are in fact positive.

The Cochrane reviews give much more weight to statistical evidence from randomised clinical trials than to prior evidence based on mechanism. It is argued that the strong emphasis on statistical evidence promotes objectivity. However, the wisdom of this approach, is questioned by some philosophers of science, Bayesian statisticians and some medical researchers [10, 11]. This paper shows that many negative Cochrane reviews are probably wrong. The opposite problem of undue enthusiasm for treatments has received more attention [12, 13]. However, this paper using a limited amount of additional but objective prior information shows the problem of false negativity may be as important. If using objective prior probabilities of effectiveness, shows that Cochrane not infrequently draws wrong conclusions, then there is a case for incorporation of more prior knowledge. In particular, assessment of the effectiveness of treatments in medicine may be improved by dropping a rigid adherence to the criterion of objectivity which prevents the cautious use of knowledge of mechanism and expert opinion as a component of prior probability.

Competing interests and acknowledgements: no competing interests; I thank anonymous reviewers for their helpful comments.

REFERENCES

1 Kault D, Kault S. From p-Values to Objective Probabilities in Assessing Medical Treatments. PLOS ONE 24/11/2015 http://journals.plos.org/plosone/article?id=10.1371/j

2 Cochrane Library http://onlinelibrary.wiley.com/cochranelibrary/search. Accessed first quarter 2016

3 Dowden J. Conflict of interest in medical journals. Australian Prescriber 2015; 38(1), pp2-3

4 Goldacre B. Bad Pharma: How drug companies mislead doctors and harm patients. Fourth Estate, 2012 (UK).

5 Seife C. Research Misconduct Identified by the US Food and Drug Administration. Out of Sight, Out of Mind, Out of the Peer-Reviewed Literature. JAMA Internal Medicine. online 9/2/2015 (accessed 29/03/2015)

6 Dubben H, Beck-Bornholdt H. Systematic review of publication bias in studies on publication bias. BMJ 2005;331:433

7 Bishop YM, Feinberg SE, Holland PW. Chapter 14, Discrete Multivariate Analysis. MIT Press 1975

8 Song F, Parekh-Bhurke S, Hooper L, et al. Extent of publication bias in different categories of research cohorts: a meta-analysis of empirical studies. BMC Medical Research Methodology 2009; 19:79 DOI: 10.1186/1471-2288-9-79

9 Kault D. False negatives in evidence based medicine. Journal of Medical Statistics and Informatics. July 2014 DOI: http://dx.doi.org/10.7243/2053-7662-2-5

10 Worrall J. Evidence in Medicine and Evidence-Based Medicine. Philosophy Compass (2007) 2/6 981-1002

11 Atwood K.Of SBM and EBM Redux. Part I: Does EBM Undervalue Basic Science and Overvalue RCTs? 12/11/2010 Blog in Science-Based Medicine website. https://www.sciencebasedmedicine.org/of-sbm-and-ebm-redux-part-i-does-ebm-undervalue accessed June 2016

12 Prasad V, Cifu A. Medical Reversal: Why We Must Raise the Bar Before Adopting New Technologies. Yale J Biol Med. 2011 Dec; 84(4): 471–478.

13 Prasad V, Vandross A, Toomey C et al. A Decade of Reversal: An Analysis of 146 Contradicted Medical Practices. Mayo Clinic Proceedings, August 2013Volume 88, Issue 8, Pages 790–798 http://dx.doi.org/10.1016/j.mayocp.2013.05.012

Appendix 1 - The police deduction from the uncertain Martian observation

We assume that the Martian's observation has two components **X**, an unknown value from the range of values of adult male human heights and **Y**, an unknown value from the range of errors that could be made by the Martian. The calculation

assumes that the range of errors is normally distributed, centred on zero and the gap between 170 and 230 represents 2×1.96 standard deviations. The standard deviation of the errors in the Martian's observation is therefore $\frac{230-170}{2\times 1.96} = 15.3$ In symbols then we have $\mathbf{Y} \sim N(0, s^2) = N(0, 15.3^2)$. We will assume that the size of the error that could be made by the Martian will be independent of the assailant's actual height. We will also assume that the police know that in their country, the distribution of human male heights is given by $\mathbf{X} \sim N(\mu, \sigma^2) = N(170, 10.2^2)$ so that 95% are between 150cm and 190cm. The outcome that the Martian could possibly observe, is $\mathbf{U} = \mathbf{X} + \mathbf{Y}$. As the sum of two independent normal variables $\mathbf{U} = \mathbf{X} + \mathbf{Y} \sim N(\mu, \sigma^2 + s^2) = N(170, 15.3^2 + 10.2^2)$. We want to find the probability distribution of the height of the man $\mathbf{X}$ given knowledge of the observation $\mathbf{U}$. We denote this probability distribution as $f_{\mathbf{X}|\mathbf{U}} = \frac{f_{\mathbf{X},\mathbf{U}}}{f_{\mathbf{U}}}$. To find $f_{\mathbf{X},\mathbf{U}}$ we note that $f_{\mathbf{X},\mathbf{U}} = f_{\mathbf{X},\mathbf{Y}}|\mathbf{J}|$ where $\mathbf{J}$ is the Jacobian $= 1$. We also note that by independence $f_{\mathbf{X},\mathbf{Y}} = f_{\mathbf{X}} f_{\mathbf{Y}} = f_{\mathbf{X}}(x) f_{\mathbf{Y}}(u - x)$. We then have:-

$$f_{\mathbf{X}|\mathbf{U}} = \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \frac{1}{\sqrt{2\pi}s} e^{\frac{-1}{2}\left(\frac{u-x}{s}\right)^2}}{\frac{1}{\sqrt{2\pi(\sigma^2+s^2)}} e^{\frac{-1}{2}\left(\frac{u-\mu}{\sqrt{\sigma^2+s^2}}\right)^2}}$$

After some algebra we find that $f_{\mathbf{X}|\mathbf{U}} \sim N(\mu_{\mathbf{X}|\mathbf{U}}, \sigma^2_{\mathbf{X}|\mathbf{U}})$ where

$$\sigma_{\mathbf{X}|\mathbf{U}} = \sqrt{\frac{\sigma^2 s^2}{\sigma^2 + s^2}} \tag{1}$$

and

$$\mu_{\mathbf{X}|\mathbf{U}} = \frac{\mu s^2 + u\sigma^2}{\sigma^2 + s^2} \tag{2}$$

Using the values of μ, σ and s given above, and taking the point estimate of u to be 200cm, we have $\mu_{\mathbf{X}|\mathbf{U}} = 179.2$cm and $\sigma_{\mathbf{X}|\mathbf{U}} = 8.5$cm. The 95% credible interval $\mu_{\mathbf{X}|\mathbf{U}} \pm 1.96\sigma_{\mathbf{X}|\mathbf{U}}$ is then (162.6cm,195.9cm).

It is relevant to note that (1) can be written as:
$\sigma_{\mathbf{X}|\mathbf{U}} = \sigma\sqrt{\frac{s^2}{\sigma^2+s^2}} = s\sqrt{\frac{\sigma^2}{\sigma^2+s^2}} < \min(\sigma, s)$ so the 95% credible interval must be smaller than the original 95% confidence interval

and that (2) can be written as:
$\mu_{\mathbf{X}|\mathbf{U}} = \frac{s^2}{\sigma^2+s^2}\mu + \frac{\sigma^2}{\sigma^2+s^2}u = w_x\mu_x + w_y u$ where $w_x + w_y = 1$ So $\mu_{\mathbf{X}|\mathbf{U}}$ is a weighted average of μ and the observation u with the weights being proportional to the precision (ie. inverse variance) of $\mathbf{X}$ and of $\mathbf{Y}$

Appendix 2 - Extension of the method to credible intervals for clinical trials

We assume that the uncertainty about the true value of effectiveness of a treatment is an amount that is normally distributed, when effectiveness is measured as the log of the relative risk [7]. As in Appendix 1 we use the symbol $\mathbf{Y}$ for a value from this distribution, with the standard deviation s, determined by the width of the confidence interval of the log relative risk so that as previously $\mathbf{Y} \sim N(0, s^2)$. We use the symbol $\mathbf{X}$ for a value from the distribution of values of the true effectiveness of a treatment. This distribution is described below. For both the Martian's estimate of height and for the clinical trial's estimate of effectiveness, $\mathbf{U}$ represents the outcome that may be observed. As in Appendix 1, $\mathbf{U} = \mathbf{X} + \mathbf{Y}$ and our task involves finding $f_{\mathbf{X}|\mathbf{U}}$.

In the first instance, it is assumed that the distribution of $\mathbf{X}$, comprises a proportion at precisely zero effectiveness with the remainder chosen from a normal distribution centred on a positive number, so

$$\mathbf{X} \sim f_{\mathbf{X}}(x) = p\delta(x) + (1-p)\phi(x, \mu, \sigma^2) \tag{3}$$

where the $p\delta(x)$ term uses the Dirac delta function and represents a spike of probability at zero of size p meaning that this proportion of treatments is precisely useless and $\phi(x, \mu, \sigma^2)$ is used for the normal probability density function. There is no convincing theoretical reason and only a modest amount of empirical evidence to support this form for $\mathbf{X}$. However, the method here may not be too sensitive to the precise form of the effectiveness distribution [9] and in any case, as described below, the distribution we eventually employ in the calculations allows greater flexibility.

As in Appendix 1 we want $f_{\mathbf{X}|\mathbf{U}} = \frac{f_{\mathbf{X},\mathbf{U}}}{f_{\mathbf{U}}}$ and as before, the numerator $f_{\mathbf{X},\mathbf{U}} = f_{\mathbf{X}}(x) f_{\mathbf{Y}}(u-x)$. The numerator $f_{\mathbf{X},\mathbf{U}}$ is then

$$\begin{aligned} f_{\mathbf{X},\mathbf{U}}(x,u) &= p\phi(u-x, 0, s^2) \times \delta(x) + (1-p) \times \phi(x, \mu, \sigma^2) \times \phi(u-x, 0, s^2) \\ &= p\phi(u, 0, s^2) \times \delta(x) + (1-p) \times \phi(x, \mu, \sigma^2) \times \phi(u-x, 0, s^2) \end{aligned}$$

Now, with some algebra it can be shown that the terms on the right side of the line above satisfy

$$\phi(x, \mu, \sigma^2) \times \phi(u-x, 0, s^2) = \phi(u, \mu, \sigma^2 + s^2) \times \phi(x, \mu_*, \sigma_*^2)$$

where $\mu_* = \frac{s^2\mu + \sigma^2 u}{\sigma^2 + s^2}$ and $\sigma_* = \frac{\sigma s}{\sqrt{\sigma^2 + s^2}}$ so we have

$$f_{\mathbf{X},\mathbf{U}}(x,u) = p\phi(u, 0, s^2) \times \delta(x) + (1-p) \times \phi(u, \mu, \sigma^2 + s^2) \times \phi(x, \mu_*, \sigma_*^2)$$

We then have

$$\begin{aligned} f_{\mathbf{U}}(u) &= \int_{-\infty}^{\infty} f_{\mathbf{X},\mathbf{U}}(x,u)dx \\ &= p\phi(u,0,s^2) + (1-p) \times \phi(u,\mu,\sigma^2+s^2) \end{aligned}$$

We now obtain

$$\begin{aligned} f_{\mathbf{X}|\mathbf{U}}(x|u) = \frac{f_{\mathbf{X},\mathbf{U}}}{f_{\mathbf{U}}} &= \frac{p\phi(u,0,s^2) \times \delta(x) + (1-p) \times \phi(u,\mu,\sigma^2+s^2) \times \phi(x,\mu_*,\sigma_*^2)}{f_{\mathbf{U}}(u)} \\ &= p_*(u)\delta(x) + (1-p_*(u))\phi(x,\mu_*,\sigma_*^2) \end{aligned} \tag{4}$$

where

$$p_* = p_*(u) = \frac{p\phi(u,0,s^2)}{f_{\mathbf{U}}(u)}$$

Hence the distribution of $f_{\mathbf{X}|\mathbf{U}}(x|u)$ given in (4), like the distribution of $f_{\mathbf{X}}(x)$ itself given in (3), is a mixture distribution of a spike at zero and a normal, but with parameters that depend via the definitions of μ_* and σ_* on u and s as well as on the parameters p, μ, σ of the original $f_{\mathbf{X}}(x)$ distribution. Inspection of the expressions for μ_* and σ_* show that they are identical in form to equation (2) and (1) respectively, and so contraction of the confidence interval and a weighted shift of the mean would be expected to apply here. This density function would also be the key to defining an interval containing 95% of the probability of $\mathbf{X}$ in $f_{\mathbf{X}|\mathbf{U}}(x|u)$, if we were certain about the parameters p, μ and σ. However, unlike the police in Appendix 1, who have virtually certain knowledge of the parameters of the normal distribution describing the height of humans based on measuring thousands of humans with an accurate ruler, we have only uncertain knowledge of the distribution of effectiveness based on fitting a curve to a sample of 101 relative risks [1]. Furthermore, those 101 relative risks in Cochrane were themselves not measured with an accurate "ruler", but were measured with a blurring of uncertainty described by the confidence interval for each item of data.

To allow for this uncertainty and to allow for publication bias, appropriately counter-biased bootstrapping was performed. In particular the estimation of p, μ and σ obtained by curve fitting, was repeated with different sets of 101 relative risks. Each of these sets of 101 relative risks were obtained by resampling with replacement from the observed set of 101 relative risks. However, this resampling was adjusted to over-represent the non-significant results by a factor of 2.78 to counteract publication bias [8]. This process gives 1000 sets of the three parameters p, μ and σ. We then index these parameters by (i) so we have a 1000 sets $\{p^{(i)}, \mu^{(i)}, \sigma^{(i)}\}$ and correspondingly we have a 1000 functions $f_{\mathbf{X}}^{(i)}(x)$ Our knowledge of $f_{\mathbf{X}}(x)$ is defined to be a sort of average of all the $f_{\mathbf{X}}^{(i)}(x)$, or more precisely, a

mixture distribution in which any component has a $\frac{1}{1000}$ probability of being chosen. Denoting this "average" by $f^{(a)}_{\mathbf{X}}(x)$ we have

$$\begin{aligned} f^{(a)}_{\mathbf{X}}(x) &= \frac{1}{1000}\sum_{i=1}^{1000}\Big(p^{(i)}\delta(x) + (1-p^{(i)})\phi(x,\mu^{(i)},\sigma^{(i)2}\Big) \\ &= p^{(a)}\delta(x) + \frac{1}{1000}\sum_{i=1}^{1000}\Big((1-p^{(i)})\phi(x,\mu^{(i)},\sigma^{(i)2})\Big) \end{aligned} \tag{5}$$

where $p^{(a)}$ is the average of the $p^{(i)}$. The shape of this prior distribution $f_{\mathbf{X}}$ is shown in Figure 1 in the main text.

Noting the similarity between (3) and (4) we likewise have

$$\begin{aligned} f^{(a)}_{\mathbf{X}|\mathbf{U}}(x|u) &= \frac{1}{1000}\sum_{i=1}^{1000}\Big(p^{(i)}_*\delta(x) + (1-p^{(i)}_*)\phi(x,\mu^{(i)}_*,\sigma^{(i)2}_*\Big) \\ &= p^{(a)}_*\delta(x) + (1-p^{(a)}_*)\frac{1}{1000}\sum_{i=1}^{1000}\Big(\frac{(1-p^{(i)}_*)}{(1-p^{(a)}_*)}\phi(x,\mu^{(i)}_*,\sigma^{(i)2}_*)\Big) \end{aligned} \tag{6}$$

where $p^{(a)}_*$ is the average of the $p^{(i)}_*$

We see that $f^{(a)}_{\mathbf{X}|\mathbf{U}}(x|u)$ retains a simple spike at zero. However the other component is not as neat. It is a mixture distribution of normals and generally will not be accurately approximated by a single normal nor will it even be symmetric.

It will be convenient to write the probability density function of this mixture distribution of normals as $\beta(x,\{p_{u_i},\mu_{u_i},\sigma_{u_i}\})$ to give the form of $f_{\mathbf{X}|\mathbf{U}}(x|u)$ we require for use in what follows:

$$f^{(a)}_{\mathbf{X}|\mathbf{U}}(x|u) = p^{(a)}_*\delta(x) + (1-p^{(a)}_*)\beta(x,\{p^{(i)}_*,\mu^{(i)}_*,\sigma^{(i)}_*\})$$

With this probability density function for the amount of true effectiveness $\mathbf{X}$ given an observation of a particular value of effectiveness u of $\mathbf{U}$ together with the measurement of uncertainty of this observation s, it is now straightforward to develop an algorithm to find the shortest interval in the distribution of $\mathbf{X}$ that contains 95% of the probability - the 95% credible interval. However, we note that the added complexity of the problem here means that although there are equations involved similar to (1) and (2), the contraction of the distribution and the shift of its centre, are not straightforward.

In the algorithm to define the 95% credible interval, there are three main cases. If $p^{(a)}_* > 0.95$ the credible interval is of length zero and located at zero. If $0.05 < p^{(a)}_* < 0.95$ the credible interval must contain the spike at zero, but there are three sub-options. The spike can be internal in the credible interval or on left or right boundary (the spike at zero on the right boundary of the interval

though would likely only arise when the treatment was quite counterproductive). If $p_*^{(a)} < 0.05$ there is the additional sub-option that the credible interval may not include the spike at zero - this spike will be very small when a convincing result is obtained. To decide on which of these sub-options might apply we use a search procedure to find the shortest interval (x_{l_1}, x_{u_1}) such that $(1 - p_*^{(a)}) \int_{x_{l_1}}^{x_{u_1}} \beta(x, \{p_*^{(i)}, \mu_*^{(i)}, \sigma_*^{(i)}\})dx = 0.95$ and the shortest interval (x_{l_2}, x_{u_2}) such that $(1-p_*^{(a)}) \int_{x_{l_2}}^{x_{u_2}} \beta(x, \{p_*^{(i)}, \mu_*^{(i)}, \sigma_*^{(i)}\})dx = 0.95 - p_*^{(a)}$. We refer to these intervals as interval 1 and interval 2 respectively. Some thought shows that if both these intervals contain the spike at zero, we have an internal spike and interval 2 is the credible interval. If neither contain the spike, interval 1 is the credible interval. If interval 1 contains the spike, but interval 2 doesn't, then the spike is on a boundary and (after a check to identify which boundary - usually the left - assumed in what follows) a search procedure is performed to find x_{u_3} such that $(1 - p_*^{(a)}) \int_0^{x_{u_3}} \beta(x, \{p_*^{(i)}, \mu_*^{(i)}, \sigma_*^{(i)}\})dx = 0.95 - p_*^{(a)}$. The credible interval is then $[0, x_{u_3})$.

To calculate the expected relative risk, E(RR) we multiply equation (3) by e^x and integrate to give

$$\begin{aligned}
\int_{-\infty}^{\infty} e^x f_{\mathbf{X}|\mathbf{U}}^{(a)}(x|u)\mathrm{d}x &= \int_{-\infty}^{\infty} e^x \frac{1}{1000} \sum_{i=1}^{1000} \Big(p_*^{(i)} \delta(x) + (1 - p_*^{(i)}) \phi(x, \mu_*^{(i)}, \sigma_*^{(i)2}) \Big) \mathrm{d}x \\
&= p_*^{(a)} + \frac{1}{1000} \sum_{i=1}^{1000} (1 - p_*^{(i)}) \int_{-\infty}^{\infty} e^x \phi(x, \mu_*^{(i)}, \sigma_*^{(i)2}) \mathrm{d}x \\
&= p_*^{(a)} + \frac{1}{1000} \sum_{i=1}^{1000} (1 - p_*^{(i)}) \Big(e^{\mu_*^{(i)} + \frac{1}{2}\sigma_*^{(i)2}} \Big) \qquad (7)
\end{aligned}$$

The work described here could be improved by using further data for calibration and perhaps calibration for particular branches of medicine and for new drugs where the data emerge from pre-registered trials. Calibration could also be specialised to particular types of statistics with perhaps adjustment for the approximations used in confidence interval calculations. Continuous data could be used if, instead of the p-value, effect size was measured in comparison to the size of the background variability. Whilst an effect of fitting one alternative model was explored elsewhere [9], much more extensive testing of the sensitivity of the results to the model used for prior distribution is also desirable. However, all this is beyond the scope of this paper.

The program to code for the calculations will be made available on James Cook University's Tropical Data Hub and then on Research Data Australia https://researchdata.ands.org.a

Table 1. Numbered List of Cochrane Studies Used as Data

Study Number	**Study Title** (truncated if necessary at around 90 characters)	**Statistic Used and effect direction**
1	Routine pre-pregnancy health promotion for improving pregnancy outcomes	RR 1.24 (1.06, 1.44) +ve
2	Routine surgery in addition to chemotherapy for treating spinal tuberculosis	OR 0.88 (0.36, 2.16) +ve
3	Home-based versus centre-based cardiac rehabilitation	RR 0.79 (0.43, 1.47) +ve
4	Interventions for pain with intrauterine device insertion	OR 0.02 (0.01, 0.09) +ve
5	Oral immunoglobulin for preventing necrotizing enterocolitis in preterm and low birth weight	RR 0.84 (0.57, 1.25) +ve
6	Topical capsaicin (low concentration) for chronic neuropathic pain in adults	RR 2.04 (0.75, 5.54) +ve
7	Day hospital versus outpatient care for people with schizophrenia	RR 0.71 (0.56, 0.89) +ve
8	Steroids for improving recovery following tonsillectomy in children	RR 0.49 (0.41, 0.58) +ve
9	Decompressive surgery for treating nerve damage in leprosy	RR 1.13 (0.10, 1.77) +ve
10	On-site mental health workers delivering psychological therapy and psychosocial interventions	RR 0.67 (0.56, 0.79) +ve
11	Enteral nutrition for maintenance of remission in Crohn's disease	OR 0.30 (0.09, 0.94) +ve
12	Dietary calcium supplementation for preventing colorectal cancer and adenomatous polyps	OR 0.74 (0.58, 0.95) +ve
13	Low molecular weight heparin for prevention of microvascular occlusion in digital replantation	RR 1.03 (0.87, 1.22) -ve
14	Intercessory prayer for the alleviation of ill health	RR 1.00 (0.74, 1.36) +/-
15	Alpha-2 adrenergic agonists for the prevention of cardiac complications among patients undergoing surgery	RR 0.66 (0.44, 0.98) +ve
16	Interventions for prevention of post-operative recurrence of Crohn's disease	RR 0.23 (0.09, 0.57) +ve
17	Open Mesh versus non-Mesh for groin hernia repair	OR 0.37 (0.26, 0.51) +ve

Continued on next page

Table 1 – continued from previous page

Study Number	Study Title	Statistic Used and effect direction
18	Pelvic floor muscle training for prevention and treatment of urinary and faecal incontinence in antenatal	RR 0.74 (0.58, 0.94) +ve
19	Chest shielding for prevention of a haemodynamically symptomatic patent ductus arteriosus in preterm	RR 0.23 (0.05, 1.01) +ve
20	Antithrombotic therapy for improving maternal or infant health outcomes in women considered at risk	RR 0.40 (0.20, 0.78) +ve
21	Methotrexate for ankylosing spondylitis	RR 3.18 (1.03, 9.79) +ve
22	External cephalic version for breech presentation before term	RR 1.04 (0.64, 1.69) -ve
23	Early versus late initiation of epidural analgesia for labour	RR 1.02 (0.96, 1.08) +ve
24	Corticosteroids for the resolution of malignant bowel obstruction in advanced gynaecological	OR 0.51 (0.21,1.23) +ve
25	Antibiotic regimens for suspected early neonatal sepsis	RR 0.75 (0.19, 2.90) -ve
26	Patient education in the management of coronary heart disease	RR 0.79 (0.55,1.13) +ve
27	Cotrimoxazole prophylaxis for opportunistic infections in adults with HIV	RR 0.69 (0.55, 0.87) +ve
28	Erythromycin for the prevention of chronic lung disease in intubated preterm infants	RR 1.40 (0.72, 2.70) -ve
29	Vitamin D supplementation in pregnancy	RR 0.52 (0.25,1.05) +ve
30	Thrombelastography (TEG) or thromboelastometry (ROTEM) to monitor haemotherapy versus usual care	RR 0.77 (0.35,1.72) +ve
31	Unit-dose packaged drugs for treating malaria	RR 1.18 (1.12,1.25) +ve
32	Therapy-based rehabilitation services for patients living at home more than one year after stroke	RR 0.85 (0.22, 3.26) +ve
33	Lecithin for dementia and cognitive impairment	OR 3.01 (0.92, 9.81) -ve
34	Antiviral treatment for preventing postherpetic neuralgia	RR 0.75 (0.51, 1.11) +ve
35	Ethamsylate for the prevention of morbidity and mortality in preterm	RR 0.63 (0.47, 0.86) +ve
36	Intensive glucose control versus conventional glucose control for type 1 diabetes mellitus	RR 0.27 (0.18, 0.42) +ve

Continued on next page

Table 1 – continued from previous page

Study Number	Study Title	Statistic Used and effect direction
37	Palliative endobronchial brachytherapy for non-small cell lung cancer	RR 1.57 (0.22, 11.08) -ve
38	Ursodeoxycholic acid for primary biliary cirrhosis	RR 0.97 (0.67, 1.42) +ve
39	Non-steroidal anti-inflammatory drugs (NSAIDs) for treating lateral elbow pain in adults	RR 1.49 (1.04, 2.14) +ve
40	Transfusion thresholds and other strategies for guiding allogeneic red blood cell transfusion	RR 0.77 (0.62, 0.95) +ve
41	Competitions and incentives for smoking cessation	OR 1.42 (1.19, 1.69) +ve
42	Cranial irradiation for preventing brain metastases of small cell lung cancer	RR 0.84 (0.73, 0.97) +ve
43	Stem cell treatment for acute myocardial infarction	RR 0.93 (o.58, 1.50) +ve
44	Single layer versus double layer suture anastomosis of the gastrointestinal tract	OR 0.76 (0.44,1.32) -ve
45	Interventions for tubal ectopic pregnancy	OR 0.28 (0.09, 0.86) +ve
46	Intravitreal low molecular weight heparin and 5-Fluorouracil for the prevention of proliferative	RR 0.48 (0.25, 0.92) +ve
47	Amphetamines for improving recovery after stroke	OR 1.45 (0.64, 3.27) -ve
48	Perineal techniques during the second stage of labour for reducing perineal trauma	RR 0.48 (0.28, 0.84) +ve
49	Interventions for preventing delirium in older people in institutional long-term care	RR 0.85 (0.18, 4.00) +ve
50	Interventions for treating sexual dysfunction in patients with chronic kidney disease	RR 1.00 (0.08, 13.02) +/-
51	Bone morphogenetic protein (BMP) for fracture healing in adults	RR 1.19 (0.99, 1.43) +ve
52	Loxapine for schizophrenia	RR 0.30 (0.14, 0.63) +ve
53	Co-bedding in neonatal nursery for promoting growth and neurodevelopment in stable preterm twins	RR 0.85 (0.18, 4.05) +ve
54	Zinc supplements for preventing otitis media	R1 0.69 (0.61, 0.79) +ve
55	Lidocaine for preventing postoperative sore throat	RR 0.64 (0.48, 0.85) +ve
56	Early versus delayed initiation of continuous distending pressure for respiratory distress	RR 0.55 (0.32, 0.96) +ve

Continued on next page

Table 1 – continued from previous page

Study Number	Study Title	Statistic Used and effect direction
57	Pre-operative radiotherapy and curative surgery for the management of localized rectal carcinoma	OR 0.93 (0.87, 1.00) +ve
58	Fibrin glue instillation under skin flaps to prevent seroma-related morbidity following breast	RR 1.02 (0.90, 1.16) -ve
59	Brief co-incubation of sperm and oocytes for in vitro fertilization techniques	OR 2.42 (1.55, 3.77) +ve
60	Sodium bicarbonate infusion during resuscitation of infants at birth	RR 1.04 (0.49,2.21) +ve
61	High feedback versus low feedback of prenatal ultrasound for reducing maternal anxiety	RR 3.30 (0.73, 14.85) +ve
62	Propranolol for migraine prophylaxis	RR 1.72 (1.23, 2.40) +ve
63	Bispectral index for improving anaesthetic delivery and postoperative recovery	OR 0.24 (0.12, 0.48) +ve
64	Vitamin K antagonists versus antiplatelet therapy after transient ischaemic attack	RR 0.80 (0.56, 1.14) +ve
65	Motivational interviewing for improving outcomes in youth living with HIV	OR 1.23 (0.57, 2.63) -ve
66	Substitution of doctors by nurses in primary care	RR 1.02 (0.98, 1.05) -ve
67	Regular treatment with salmeterol and inhaled steroids for chronic asthma: serious adverse events	OR 0.90 (0.31, 2.60) -ve
68	Isotonic versus hypotonic solutions for maintenance intravenous fluid administration in children	RR 0.48 (0.38, 0.60) +ve
69	Intermittent pneumatic compression for treating venous leg ulcers	RR 2.27 (1.30, 3.97) +ve
70	Intravenous oxytocin alone for cervical ripening and induction of labour	RR 0.16 (0.10, 0.25) +ve
71	Topical NSAIDs for chronic musculoskeletal pain in adults	RR 1.17 (1.08, 1.26) +ve
72	Endoscopic third ventriculostomy (ETV) for idiopathic normal pressure hydrocephalus (iNPH)	OR 1.12 (0.26, 4.76) +ve
73	Interventions for preventing falls in people after stroke	R1 0.92 (0.45, 1.90) +ve

Continued on next page

Table 1 – continued from previous page

Study Number	Study Title	Statistic Used and effect direction
74	Antioxidant supplements for preventing gastrointestinal cancers	RR 0.94 (0.83 1.06) +ve
75	Mucolytic agents for chronic bronchitis or chronic obstructive pulmonary disease	OR 1.75 (1.57, 1.94) +ve
76	Interventions for preventing infection in nephrotic syndrome	RR 0.47 (0.31,0.73) +ve
77	Hydration for treatment of preterm labour	RR 1.09 (0.71,1.68) -ve
78	Grey literature in meta-analyses of randomized trials of health care interventions	R2 1.09 (1.03, 1.16) +ve
79	Heparin versus 0.9% sodium chloride intermittent flushing for prevention of occlusion in central venous	RR 0.53 (0.29, 0.94) +ve
80	Herbal medicines for fatty liver diseases	RR 0.36 (0.14, 0.97) +ve
81	Patient education for preventing diabetic foot ulceration	RR 0.31 (0.14, 0.65) +ve
82	Budesonide versus placebo for chronic asthma in children and adults	RR 0.17 (0.09, 0.33) +ve
83	Combined use of hyperthermia and radiation therapy for treating locally advanced cervix carcinoma	RR 0.56 (0.39, 0.79) +ve
84	Active cycle of breathing technique for cystic fibrosis	RR 1.64 (0.62, 4.34) +ve
85	Opioid antagonists for smoking cessation	RR 1.00 (0.66, 1.51) +/-
86	Ergonomic design and training for preventing work-related musculoskeletal disorders of the upper limb	RR 0.52 (0.27, 0.99) +ve
87	Antibiotics prior to embryo transfer in ART	OR 1.02 (0.66, 1.58) -ve
88	Vaccines for preventing smallpox	RR 0.97 (0.90, 1.03) +ve
89	Testosterone for schizophrenia	RR 0.23 (0.09, 0.57) +ve
90	Optimal monitoring strategies for guiding when to switch first-line antiretroviral therapy regimens	HR 1.35 (1.12, 1.63) +ve
91	Excisional surgery versus ablative surgery for ovarian endometriomata	OR 0.15 (0.06, 0.38) +ve
92	Enteral nutrition formulations for acute pancreatitis	RR 0.49 (0.29, 0.80) +ve
93	Colchicine for acute gout	RR 2.16 (1.28, 3.65) +ve

Continued on next page

Table 1 – continued from previous page

Study Number	Study Title	Statistic Used and effect direction
94	Tricyclic antidepressants for attention deficit hyperactivity disorder (ADHD) in children and adolescents	OR 18.50 (6.29, 54.39) +ve
95	Artificial and bioartificial support systems for liver failure	RR 0.86 (0.65, 1.14) +ve
96	Slum upgrading strategies involving physical environment and infrastructure interventions	RR 0.53 (0.27, 1.04) +ve
97	Biocompatible hemodialysis membranes for acute renal failure	RR 0.93 (0.81, 1.07) +ve
98	Cognitive behavioural therapy for men who physically abuse their female partner	RR 0.86 (0.54, 1.38) +ve
99	Abdominal drainage to prevent intra-peritoneal abscess after open appendectomy	RR 1.23 (0.47, 3.21) -ve
100	Lidocaine for reducing propofol -induced pain on induction of anaesthesia in adults	OR 0.19 (0.15,0.25) +ve
Notes	RR - relative risk OR - Odds Ratio HR - Hazard Ratio R1 - rate ratio R2 - ratio of Odds ratios	

Table 2. Results for each study

Abbreviated Column Headings*									
#	RR	Orig 95% conf int	Dirn Expctd	Prob Effect < 0	$=0$	> 0	CCI	E(RR)	Dirn Expctd	Rank
1	1.24	(1.06, 1.44)	Yes	0.00	0.07	0.93	[1.00, 1.41)	1.24	Yes	3
2	0.88	(0.36, 2.16)	Yes	0.05	0.41	0.54	(0.48, 1.11)	0.81	Yes	4
3	0.79	(0.43, 1.47)	Yes	0.04	0.38	0.58	(0.51, 1.06)	0.81	Yes	3
4	0.02	(0.01, 0.09)	Yes	0.00	0.00	1.00	(0.12, 0.84)	0.26	Yes	4
5	0.84	(0.57, 1.25)	Yes	0.04	0.44	0.53	(0.60, 1.03)	0.86	Yes	2
6	2.04	(0.75, 5.54)	Yes	0.02	0.22	0.76	(0.98, 2.73)	1.51	Yes	2
7	0.71	(0.56, 0.89)	Yes	0.00	0.03	0.97	(0.55, 0.91)	0.71	Yes	4
8	0.49	(0.41, 0.58)	Yes	0.00	0.00	1.00	(0.42, 0.62)	0.51	Yes	5
9	1.13	(0.71, 1.77)	Yes	0.05	0.48	0.47	(0.94, 1.66)	1.14	Yes	3
10	0.67	(0.56, 0.79)	Yes	0.00	0.00	1.00	(0.57, 0.80)	0.67	Yes	4

Continued on next page

Table 2 – continued from previous page

Abbreviated Column Headings*									
#	RR	Orig 95% conf int	Dirn Expctd	Prob Effect <0	$=0$	>0	CCI	E(RR)	Dirn Expctd	Rank
11	0.30	(0.09, 0.94)	Yes	0.02	0.14	0.84	(0.31, 1.00]	0.58	Yes	4
12	0.74	(0.58, 0.95)	Yes	0.00	0.10	0.90	(0.61, 1.00]	0.75	Yes	4
13	1.03	(0.87, 1.22)	No	0.10	0.76	0.14	(0.88, 1.09)	0.99	Yes	3
14	1.00	(0.74, 1.36)	+/-	0.08	0.64	0.28	(0.75, 1.09)	0.96	Yes	2
15	0.66	(0.44, 0.98)	Yes	0.01	0.13	0.87	(0.50, 1.00]	0.70	Yes	4
16	0.23	(0.09, 0.57)	Yes	0.00	0.03	0.97	(0.23, 0.95)	0.48	Yes	5
17	0.37	(0.26, 0.51)	Yes	0.00	0.00	1.00	(0.31, 0.68)	0.44	Yes	5
18	0.71	(0.54, 0.95)	Yes	0.00	0.09	0.91	(0.57, 1.00]	0.72	Yes	5
19	0.23	(0.05, 1.01)	Yes	0.02	0.17	0.81	(0.31, 1.02)	0.60	Yes	3
20	0.40	(0.20, 0.78)	Yes	0.00	0.04	0.95	(0.34, 1.00]	0.54	Yes	4
21	3.18	(1.03, 9.79)	Yes	0.01	0.14	0.85	[1.00, 3.16)	1.72	Yes	3
22	1.04	(0.64, 1.69)	No	0.08	0.57	0.35	(0.66, 1.14)	0.93	Yes	4
23	1.02	(0.96, 1.08)	Yes	0.03	0.84	0.13	[1.00, 1.06)	1.00	Yes	3
24	0.51	(0.21, 1.23)	Yes	0.02	0.20	0.78	(0.37, 1.00]	0.66	Yes	5
25	0.75	(0.19, 2.90)	No	0.07	0.43	0.49	(0.84, 2.10)	1.20	Yes	3
26	0.79	(0.55, 1.13)	Yes	0.02	0.35	0.63	(0.59, 1.00]	0.83	Yes	4
27	0.69	(0.55, 0.87)	Yes	0.00	0.01	0.99	(0.55, 0.87)	0.69	Yes	5
28	1.40	(0.72, 2.70)	No	0.13	0.58	0.29	(0.66, 1.27)	0.95	Yes	3
29	0.52	(0.25, 1.05)	Yes	0.01	0.15	0.84	(0.39, 1.00]	0.64	Yes	4
30	0.77	(0.35, 1.72)	Yes	0.04	0.36	0.59	(0.47, 1.07)	0.79	Yes	4
31	1.18	(1.12, 1.25)	Yes	0.00	0.00	1.00	(1.12, 1.25)	1.18	Yes	4
32	0.85	(0.22, 3.26)	Yes	0.05	0.37	0.58	(0.43, 1.13)	0.77	Yes	4
33	3.01	(0.92, 9.81)	No	0.14	0.54	0.32	(0.60, 1.36)	0.94	Yes	2
34	0.75	(0.51, 1.11)	Yes	0.02	0.28	0.70	(0.56, 1.00]	0.79	Yes	1
35	0.63	(0.47, 0.86)	Yes	0.00	0.02	0.98	(0.48, 0.87)	0.65	Yes	1
36	0.27	(0.18, 0.42)	Yes	0.00	0.00	1.00	(0.24, 0.68)	0.37	Yes	5
37	1.57	(0.22,11.08)	No	0.07	0.40	0.53	(0.44, 1.18)	0.80	Yes	4
38	0.97	(0.67, 1.42)	Yes	0.07	0.58	0.35	(0.69, 1.09)	0.93	Yes	4
39	1.49	(1.04, 2.14)	Yes	0.00	0.11	0.89	[1.00, 1.92)	1.43	Yes	4
40	0.77	(0.62, 0.95)	Yes	0.00	0.11	0.88	(0.65, 1.00]	0.78	Yes	5
41	1.42	(1.19, 1.69)	Yes	0.00	0.00	1.00	(1.20, 1.68)	1.42	Yes	5
42	0.84	(0.73, 0.97)	Yes	0.00	0.17	0.83	(0.74, 1.00]	0.85	Yes	5
43	0.93	(0.58, 1.50)	Yes	0.06	0.51	0.43	(0.62, 1.09)	0.90	Yes	3
44	0.76	(0.44, 1.32)	No	0.14	0.60	0.26	(0.79, 1.42)	1.03	Yes	3
45	0.28	(0.09, 0.86)	Yes	0.01	0.11	0.87	(0.31, 1.00]	0.56	Yes	2
46	0.48	(0.25, 0.92)	Yes	0.01	0.08	0.91	(0.38, 1.00]	0.59	Yes	3
47	1.45	(0.64, 3.27)	No	0.11	0.54	0.35	(0.61, 1.25)	0.92	Yes	3

Continued on next page

Table 2 – continued from previous page

Abbreviated Column Headings*									
#	RR	Orig 95% conf int	Dirn Expctd	Prob Effect <0	$=0$	>0	CCI	E(RR)	Dirn Expctd	Rank
48	0.48	(0.28, 0.84)	Yes	0.00	0.04	0.96	(0.31, 0.97)	0.57	Yes	5
49	0.85	(0.18, 4.00)	Yes	0.05	0.36	0.58	(0.42, 1.14)	0.77	Yes	3
50	1.00	(0.08,13.02)	+/-	0.06	0.36	0.58	(0.40, 1.15)	0.75	Yes	4
51	1.19	(0.99, 1.43)	Yes	0.01	0.31	0.68	[1.00, 1.39)	1.15	Yes	4
52	0.30	(0.14, 0.63)	Yes	0.00	0.02	0.98	(0.26, 0.90)	0.48	Yes	5
53	0.85	(0.18, 4.05)	Yes	0.05	0.36	0.58	(0.42, 1.14)	0.77	Yes	3
54	0.69	(0.61, 0.79)	Yes	0.00	0.00	1.00	(0.61, 0.79)	0.69	Yes	4
55	0.64	(0.48, 0.85)	Yes	0.00	0.01	0.98	(0.49, 0.86)	0.65	Yes	4
56	0.55	(0.32, 0.96)	Yes	0.01	0.09	0.90	(0.42, 1.00]	0.63	Yes	5
57	0.93	(0.87, 1.00)	Yes	0.01	0.49	0.51	(0.88, 1.00]	0.96	Yes	5
58	1.02	(0.90, 1.16)	No	0.09	0.80	0.11	(0.92, 1.06)	1.00	Yes	3
59	2.42	(1.55, 3.77)	Yes	0.00	0.00	1.00	(1.26, 3.01)	2.02	Yes	5
60	1.04	(0.49, 2.21)	No	0.07	0.48	0.44	(0.55, 1.15)	0.87	Yes	3
61	2.93	(1.25, 6.86)	Yes	0.01	0.07	0.93	[1.00, 3.12)	1.86	Yes	3
62	1.72	(1.23, 2.40)	Yes	0.00	0.01	0.99	(1.19, 2.25)	1.64	Yes	5
63	0.24	(0.12, 0.48)	Yes	0.00	0.00	1.00	(0.25, 0.83)	0.42	Yes	5
64	0.80	(0.56, 1.14)	Yes	0.02	0.36	0.61	(0.59, 1.00]	0.84	Yes	1
65	1.23	(0.57, 2.63)	No	0.09	0.53	0.38	(0.59, 1.20)	0.91	Yes	5
66	1.02	(0.98, 1.05)	No	0.10	0.88	0.02	[1.00, 1.03)	1.00	No	3
67	0.90	(0.31, 2.60)	No	0.07	0.44	0.49	(0.86, 2.02)	1.19	Yes	3
68	0.48	(0.38, 0.60)	Yes	0.00	0.00	1.00	(0.40, 0.67)	0.51	Yes	5
69	2.27	(1.30, 3.97)	Yes	0.00	0.02	0.98	(1.12, 3.09)	1.85	Yes	4
70	0.16	(0.10, 0.25)	Yes	0.00	0.00	1.00	(0.16, 0.63)	0.28	Yes	5
71	1.17	(1.08, 1.26)	Yes	0.00	0.00	1.00	(1.09, 1.27)	1.18	Yes	5
72	1.12	(0.26, 4.76)	Yes	0.06	0.37	0.57	(0.88, 2.35)	1.29	Yes	3
73	0.92	(0.45, 1.90)	Yes	0.06	0.45	0.49	(0.53, 1.11)	0.85	Yes	3
74	0.94	(0.83, 1.06)	Yes	0.03	0.68	0.29	(0.84, 1.00]	0.97	Yes	2
75	1.75	(1.57, 1.94)	Yes	0.00	0.00	1.00	(1.55, 1.94)	1.73	Yes	4
76	0.47	(0.31, 0.73)	Yes	0.00	0.00	1.00	(0.37, 0.83)	0.54	Yes	4
77	1.09	(0.71, 1.68)	No	0.10	0.61	0.29	(0.71, 1.16)	0.96	Yes	3
78	1.09	(1.03, 1.16)	Yes	0.00	0.17	0.83	[1.00, 1.15)	1.08	Yes	5
79	0.53	(0.29, 0.94)	Yes	0.01	0.10	0.90	(0.41, 1.00]	0.62	Yes	3
80	0.36	(0.14, 0.97)	Yes	0.01	0.12	0.87	(0.33, 1.00]	0.58	Yes	4
81	0.31	(0.14, 0.66)	Yes	0.00	0.03	0.97	(0.26, 0.93)	0.50	Yes	4
82	0.17	(0.09, 0.33)	Yes	0.00	0.00	1.00	(0.20, 0.77)	0.35	Yes	5
83	0.56	(0.39, 0.79)	Yes	0.00	0.01	0.99	(0.43, 0.84)	0.59	Yes	4
84	1.64	(0.62, 4.34)	Yes	0.03	0.29	0.68	(0.95, 2.48)	1.40	Yes	1

Continued on next page

Table 2 – continued from previous page

Abbreviated Column Headings*									
#	RR	Orig 95% conf int	Dirn Expctd	Prob Effect <0	Prob Effect $=0$	Prob Effect >0	CCI	E(RR)	Dirn Expctd	Rank
85	1.00	(0.66, 1.51)	+/-	0.08	0.58	0.34	(0.68, 1.11)	0.94	Yes	1
86	0.52	(0.27, 0.99)	Yes	0.01	0.12	0.87	(0.40, 1.00]	0.62	Yes	4
87	1.02	(0.66, 1.58)	No	0.08	0.58	0.34	(0.68, 1.13)	0.94	Yes	2
88	0.97	(0.90, 1.03)	Yes	0.03	0.80	0.17	(0.93, 1.00]	0.99	Yes	3
89	0.23	(0.09, 0.57)	Yes	0.00	0.03	0.97	(0.23, 0.95)	0.48	Yes	3
90	1.35	(1.12, 1.63)	Yes	0.00	0.02	0.98	(1.12, 1.65)	1.35	Yes	5
91	0.15	(0.06, 0.38)	Yes	0.00	0.01	0.99	(0.21, 0.86)	0.41	Yes	5
92	0.49	(0.29, 0.80)	Yes	0.00	0.03	0.97	(0.34, 0.91)	0.57	Yes	4
93	2.16	(1.28, 3.65)	Yes	0.00	0.02	0.98	(1.12, 2.96)	1.82	Yes	5
94	18.50	(6.29,54.39)	Yes	0.00	0.00	1.00	(1.19, 6.06)	2.95	Yes	5
95	0.86	(0.65, 1.14)	Yes	0.03	0.47	0.50	(0.67, 1.01)	0.89	Yes	4
96	0.53	(0.27, 1.04)	Yes	0.01	0.14	0.85	(0.40, 1.00]	0.64	Yes	4
97	0.93	(0.81, 1.07)	Yes	0.03	0.65	0.33	(0.82, 1.00]	0.96	Yes	1
98	0.86	(0.54, 1.38)	Yes	0.04	0.45	0.50	(0.59, 1.06)	0.86	Yes	3
99	1.23	(0.47, 3.21)	No	0.08	0.48	0.44	(0.53, 1.19)	0.87	Yes	3
100	0.19	(0.15, 0.25)	Yes	0.00	0.00	1.00	(0.17, 0.43)	0.24	Yes	5

*Table Heading Explanations	
RR	Relative Risk
Orig 95% conf int	Original 95% confidence interval
Dirn Expctd	"Does the RR tend in the anticipated direction?"
Prob Effect <0	Probability the treatment effect is opposite to that anticipated
Prob Effect $=0$	Probability there is no treatment effect ie. RR=1.00
Prob Effect >0	Probability the treatment effect is in direction anticipated
CCI	Contracted Confidence Interval. Also note:-
The square bracket " ["or "] " includes the 1.00 endpoint in the confidence interval	
E(RR)	The expected value of the relative risk
Dirn Expctd	"Does the E(RR) tend in the anticipated direction?"
Rank	The effectiveness ranking interpreted from the Cochrane summary