Health Care Manage Sci (2008) 11:121–131
DOI 10.1007/s10729-007-9041-8

A new approach to outliers in meta-analysis

Rose Baker · Dan Jackson

Received: 10 July 2007 / Accepted: 21 November 2007 / Published online: 21 December 2007

Abstract The synthesis of evidence from trials and medical studies using meta-analysis is essential for Evidence Based Medicine. However, problematical outlying results often occur even under the random-effects model. We propose a model that allows a long-tailed distribution for the random effect, which removes the necessity for an arbitrary decision to include or exclude outliers. In this approach, they are included, but with a reduced weight. We also introduce a modification of the forest plot to show the downweighting of outliers. We illustrate the methodology and its usefulness by carrying out both frequentist and Bayesian meta-analyses using data sets from the Cochrane Collaboration.

Keywords Subbotin distribution · Arcsinh transformation · Score test · Importance sampling

1 Introduction

The 'meta-analysis' of clinical trials and medical studies underpins Evidence Based Medicine (EBM), by allowing the pooling of all available evidence about the effect of a treatment or intervention. Hence there is ongoing interest in improving the statistical techniques used. Outside the field of health care, meta-analysis is also now being used in education, social work and social welfare, and crime and justice.

Meta-analysis poses many problems for the analyst, a worrying one being publication bias, which has been studied by many, including the present authors [1, 2]. However, the most obvious problem encountered when trying to form a consensus of the results of several studies is that the estimates of a treatment effect often agree less well than their quoted statistical errors would imply. Different patient mixes and operational procedures must be factors here. In situations where the necessary covariates are available, it is possible to model such treatment effect variability by performing a 'meta-regression' [3]; if these are not available we assume that what is being measured is not the true value of the effect, but an effect displaced from the true one by a random amount η [4–6]. This random effect η includes all unobservable factors such as patient mix. In our experience, roughly 75% of meta-analyses require a random effect.

The distribution of the random effect has traditionally been assumed to be Gaussian. This is the default assumption made about unknown distributions in Statistics, having the merit of simplicity, and often being motivated by the central limit theorem (CLT). Here, for studies of any reasonable size, the CLT certainly implies that the effect y_i measured in the ith of n studies should be normally distributed about some mean

R. Baker (✉)
Centre for Operational Research and Applied Statistics,
University of Salford, Salford, UK
e-mail: R.D.Baker@salford.ac.uk

D. Jackson
MRC Biostatistics Unit, Institute of Public Health,
Cambridge, UK

$\mu_i = \mu + \eta_i$ with a variance σ_i^2, which is usually assumed to be known exactly. However, the CLT does not really imply anything for the distribution of the random effect η. We can only appeal to the CLT here with the vague idea that the unknown source of variation between studies might be the sum of several factors.

In practice, we fairly often find outlying studies whose results would be extremely unlikely under the conventional random effects model. It is of course sometimes difficult to be sure whether a particular study is outlying or not, as we expect some apparent outliers to occur by chance even under a normal model; the unusual does happen occasionally. See Barnett [7] for a discussion of outliers in statistical work. Sometimes outlying studies are embarrassingly not small studies, but very large ones. We may be able to eliminate such studies if the odd result follows from poor study protocols, but every statistician knows that outliers should not be rejected merely because they are outliers [8]; it may be the model rather than the data that is at fault. We therefore need a model for meta-analysis that can cope with outliers, and which can be used routinely without prior inspection of the data. In meta-analysis, the final estimate $\hat{\mu}$ of the effect of interest is a weighted mean of the n study results y_i. Use of a potentially long-tailed random effects distribution reduces the weight given to outliers, but does not eliminate them from the analysis, and requires no arbitrary decision from the analyst.

There has been work exploring the assumption of normality in the context of meta-analysis [9], and more flexible random effects models have been suggested; see Lee and Thompson [10] for example, who explore a variety of such models and provide a wide range of references concerning work in this area. In particular, it should be noted that BUGS code, such as that provided by Smith, Speigelhalter et al. [11], can easily be amended to allow more flexible random effects models in the context of Bayesian meta-analysis. In addition to this, analogous multivariate [12] and nonparametric [13] procedures have been suggested. Despite this considerable interest in alternative formulations of the random effects model, the types of long-tailed (leptokurtic) distributions developed here have not previously been used to model outliers in meta-analysis, and we show how these novel random effects distributions can be applied in both the frequentist and Bayesian frameworks. We also explore their implications for the resulting inference and investigate whether short-tailed or platykurtic distributions are ever applicable. Data sets from the Cochrane Collaboration are used to illustrate the performance of the methodology.

2 Methodology

2.1 Inference

The likelihood function is the basis of most frequentist statistical inference, and all Bayesian inference. We assume n normally-distributed observations $y_1 \ldots y_n$ of an effect μ, with known variances σ_i^2, and a random effect with pdf $g(\eta|\tau, \phi)$, where τ is a scale parameter, and ϕ a shape parameter. Four innovative candidate distributions for g in this form are presented below, where more conventional symbols in these contexts are used instead of ϕ. For example, taking the product of τ and a random variable with a t-distribution for g, ϕ could be taken as (say) the inverse of the number of degrees of freedom ν, so that g is the pdf of the normal distribution with variance τ^2 when $\phi = 0$. Otherwise the variance is $\nu\tau^2/(\nu - 2)$ if $\nu > 2$ [14]. The pdf $f(y_i|\mu, \tau, \phi)$ of observing y_i is in general

$$f(y_i|\mu, \tau, \phi) = \frac{1}{(2\pi\sigma_i^2)^{1/2}} \int_{-\infty}^{\infty} \exp\{-(y_i - \mu - \eta)^2/2\sigma_i^2\} \times g(\eta|\tau, \phi)\, \mathrm{d}\eta. \quad (1)$$

If g is the normal pdf, so that $\eta \sim N[0, \tau^2]$ and the conventional random effects model is adopted, the integral can be evaluated analytically to give

$$f(y_i|\mu, \tau) = (2\pi(\sigma_i^2 + \tau^2))^{-1/2} \exp\left(-\frac{(y_i - \mu)^2}{2(\sigma_i^2 + \tau^2)}\right).$$

In general, the likelihood function is

$$\mathcal{L} = \prod_{i=1}^{n} f(y_i|\mu, \tau, \phi)$$

and the corresponding log-likelihood function is

$$\ell = \sum_{i=1}^{n} \ln f(y_i|\mu, \tau, \phi). \quad (2)$$

For frequentist inference we maximise ℓ to obtain maximum-likelihood estimates $\hat{\mu}, \hat{\tau}$ and $\hat{\phi}$, of which only $\hat{\mu}$ is of direct interest. Under the null hypothesis, $H_0 : \phi = 0$, twice the increase in log-likelihood is asymptotically distributed as a chi-squared with 1 degree of freedom, i.e.

$$2(\sup \ell(\mu, \tau, \phi) - \sup \ell(\mu, \tau, 0)) \sim \chi^2[1].$$

One can therefore test whether the long-tailed distribution is needed to describe the data.

When the number of studies is large (which is often not the case) a 95% confidence interval for μ can

be accurately obtained from the estimated covariance matrix of the three parameter estimates. This is obtained from the Hessian, the matrix of second derivatives $H_{ij} = -\partial^2\ell/\partial\theta_i\partial\theta_j|_{\theta=\hat{\theta}}$, where now the three model parameters μ, τ and ϕ are denoted by $\theta_1, \theta_2, \theta_3$. The estimated covariance matrix is $\mathbf{V} = \mathbf{H}^{-1}$, and the 95% confidence interval for μ is $\hat{\mu} \pm 1.96\sqrt{V_{11}}$. We prefer to compute the confidence interval from the profile likelihood, because this gives a more accurate interval, although it requires more computation. The limits of the interval are the values of μ that decrease 2ℓ from its maximum by $(1.96)^2$.

Although individual studies may be large, if there are only a few of them the sample is in a sense small, and the asymptotic properties of ℓ may not hold at all accurately. We may also feel that it is asking too much of the data to obtain meaningful estimates of three model parameters when we have only a handful of studies. We could use some method for model choice, in which for example we move from a fixed effects model ($\tau = 0$) to a conventional random effects model by introducing τ, and then to the full model, by choosing for example the minimum-AIC (Akaike Information Criterion) model [15]. This requires choosing that model as best which minimises $-2\ell + 2f$, where f is the number of model parameters. Unfortunately, the confidence interval will then be conditional on the model selected, and will be slightly too narrow. This is however probably the best frequentist approach.

For those willing to use Bayesian methods, a nice solution is to find the posterior distribution of μ, and from this to read off the mean as the 'estimate' of μ, and the 2.5 and 97.5 percentage points of the distribution as the limits of a 95% credibility interval. This posterior distribution is given by

$$f(\mu|\mathbf{y}) = \frac{\int_0^\infty \int_0^\infty \mathcal{L}(\mathbf{y}|\mu,\tau,\phi) g_\mu(\mu) g_\tau(\tau) g_\phi(\phi)\, \mathrm{d}\tau \mathrm{d}\phi}{\int_0^\infty \int_0^\infty \int_{-\infty}^\infty \mathcal{L}(\mathbf{y}|\mu,\tau,\phi) g_\mu(\mu) g_\tau(\tau) g_\phi(\phi) \mathrm{d}\tau \mathrm{d}\phi \mathrm{d}\mu}, \tag{3}$$

where g_μ, g_τ and g_ϕ are the three prior distributions for the model parameters, assumed independent. In this approach, rather than selecting a 'best' model and conditioning inference on that choice, we average over all possibilities.

Besides formal inference, plots and tables that aid understanding are useful. A new plot and a new table are suggested here. From Eq. 2, regarding τ and ϕ for the moment as known, differentiation with respect to μ to obtain the corresponding maximum likelihood estimate yields

$$\sum_{i=1}^n \partial \ln f(y_i|\mu,\tau,\phi)/\partial\mu|_{\mu=\hat{\mu}} = 0.$$

For the usual random effects model, this gives

$$\sum_{i=1}^n w_i^0 (y_i - \hat{\mu}) = 0,$$

where the weight $w_i^0 = 1/(\sigma_i^2 + \tau^2)$. This is the conventional weight. Hence we can define a weight w_i such that $\sum_{i=1}^n w_i(y_i - \hat{\mu}) = 0$, where

$$w_i = \frac{\partial \ln f(y_i|\mu,\tau,\phi)/\partial\mu|_{\mu=\hat{\mu}}}{y_i - \hat{\mu}}. \tag{4}$$

This weight reduces to w_i^0 for the normal case. Note that we could also write

$$w_i = -\frac{\partial \ln f(y_i|\mu,\tau,\phi)/\partial y_i|_{\mu=\hat{\mu}}}{y_i - \hat{\mu}}.$$

A table of the weights w_i or the ratios w_i/w_i^0 against $(y_i - \hat{\mu})$ gives an insight into the effect of the third model parameter.

Besides defining a weight w_i, we could also define an effective value of y_i as y_i^*, such that

$$y_i^* - \hat{\mu} = \frac{\partial \ln f(y_i|\mu,\tau,\phi)/\partial\mu|_{\mu=\hat{\mu}}}{w_i^0},$$

so that the weight is unchanged, but the observation y_i is adjusted. However, y_i^* may differ considerably from y_i. Visually, it seems best to simply display the ratio of weights w_i/w_i^0 against the study, and we propose that these values using the fitted models should be shown in a modified forest plot as shown in Section 3; i.e. evaluate the numerators of this ratio, w_i, using the maximum likelihood estimates $\hat{\tau}$ and $\hat{\phi}$, and similarly evaluate the denominators w_i^0 using the corresponding $\hat{\tau}$ from the usual random effects model.

From Eq. 1, we have that

$$\partial \ln f(y_i|\mu,\tau,\phi)/\partial\mu = \frac{\int_{-\infty}^\infty (y_i - \mu - \eta) \exp(-(y_i - \mu - \eta)^2/2\sigma_i^2) g(\eta)\, \mathrm{d}\eta}{\sigma_i^2 \int_{-\infty}^\infty \exp(-(y_i - \mu - \eta)^2/2\sigma_i^2) g(\eta)\, \mathrm{d}\eta}.$$

For the random effect distributions below where $g(\eta)$ is an even function, on splitting the integral at zero and expanding the exponentials, using Eq. 4 we obtain

$$w_i = 1/\sigma_i^2 - \frac{\int_0^\infty \eta \exp(-\eta^2/2\sigma_i^2)\sinh\{(y_i-\hat{\mu})\eta/\sigma_i^2\}g(\eta)\,\mathrm{d}\eta}{\sigma_i^2(y_i-\hat{\mu})\int_0^\infty \exp(-\eta^2/2\sigma_i^2)\cosh\{(y_i-\hat{\mu})\eta/\sigma_i^2\}g(\eta)\mathrm{d}\eta}.$$

Note that as $y_i \to \hat{\mu}$ the numerator and denominator tend to zero, but the expression can still be evaluated by expanding $\sinh x = x + x^3/6 + \ldots$.

2.2 Modelling

In order to model outliers, we seek long-tailed distributions that are symmetric about the (zero) mean of η, and which can depart from normality using just one extra model parameter. Four candidate distributions are described below, the latter two of which further allow for the possibility of short tailed random effect distributions.

2.2.1 The t-distribution

As mentioned above, an obvious candidate for the random effect is the product of τ and a random variable with a t-distribution, giving a pdf of

$$g(\eta|\tau,\nu) = \frac{\Gamma((\nu+1)/2)}{\tau\sqrt{\pi\nu}\Gamma(\nu/2)}(1+\eta^2/\nu\tau^2)^{-((\nu+1)/2)},$$

where $\Gamma(\cdot)$ is the gamma function, so that τ retains its interpretation as measuring the between-study variation. The t distribution has an advantage in that it can be derived by making the variance of a normal distribution a random variable with an inverse-gamma distribution. Hence it is a natural extension of the random effects model.

It is desirable to have some indication of whether a distribution is likely to be long or short tailed that can be evaluated without iterating for the third model parameter. A Rao score test [16] was devised for this purpose. The same test resulted using either the t-distribution or the arcsinh distribution (described below). A document providing the derivation of the score test statistic, using the t-distribution for the random effect, is available from either author on request and the test statistic is given by:

$$\frac{1}{2\sqrt{6}} \frac{\sum_{i=1}^n \left\{ \frac{\left(\frac{(y_i-\hat{\mu})^2}{\sigma_i^2+\hat{\tau}^2}-3\right)^2}{(\sigma_i^2+\hat{\tau}^2)^2} - \frac{6}{(\sigma_i^2+\hat{\tau}^2)^2} \right\}}{\left\{ \sum_{i=1}^n \frac{1}{(\sigma_i^2+\hat{\tau}^2)^4} \right\}^{1/2}}. \quad (5)$$

The test is not very powerful, but the sign of the statistic is useful in indicating leptokurty (positive) or platykurty (negative).

2.2.2 The arcsinh distribution

Another long-tailed distribution symmetric about the origin can be derived using Johnson's arcsinh transformation [14]. We assume that $x = \sinh^{-1}(c\eta)/c$ is normally distributed. As $c \to 0$ we regain the normal distribution. We can also write $x = \ln(c\eta + (1 + c^2\eta^2)^{1/2})/c$. The product of this random variable and τ gives the random effects density

$$g(\eta|\tau,c) = \frac{1}{(2\pi\tau^2(1+c^2\eta^2))^{1/2}} \exp\{-(\sinh^{-1}(c\eta))^2/2c^2\tau^2\}.$$

2.2.3 The beta distribution

The possibility that the distribution of η might be shorter-tailed than normal has never been considered in the meta-analysis literature. Here we would not see worrying outliers, but rather a suspicious lack of them. It is however possible that this could occur, and we consider models that allow short-tailed distributions.

A distribution family that is short-tailed, and includes the normal, was derived from the t-distribution of a variable x by making the transformation to y, where $y = x/(1+x^2/\nu)^{1/2}$. Then $\mathrm{d}y/\mathrm{d}x = (1-y^2/\nu)^{3/2}$ and $1/(1+x^2/\nu) = 1 - y^2/\nu$. The product of this random variable and τ gives the random effects density

$$g(\eta|\tau,\nu) = \frac{\Gamma((\nu+1)/2)}{\tau\sqrt{\pi\nu}\Gamma(\nu/2)}(1-\eta^2/\nu\tau^2)^{((\nu-2)/2)},$$

where $-\nu^{1/2}\tau < y < \nu^{1/2}\tau$. The relation to the t-distribution can be seen, and normality is approached as $\nu \to \infty$. This is a special case of the (scaled) beta distribution, the symmetric beta distribution [14].

2.2.4 The Subbotin distribution

A distribution that could be either long or short tailed would be attractive. These are very hard to find. The Subbotin distribution [14, 17] is one, and gives the random effects density

$$g(\eta|\tau,\delta) = \frac{1}{\tau 2^{1+\delta/2}\Gamma(1+\delta/2)} \exp\left(-\frac{1}{2}|\eta/\tau|^{2/\delta}\right).$$

For $\delta > 1$ it is long tailed, giving the Laplace distribution when $\delta = 2$, and as $\delta \to 0$ it approaches a uniform distribution with limits $\pm\tau$. Its only displeasing feature is the discontinuous first derivative of g at zero for $\delta > 1$. Note however that the function g appears in the integral in (1), resulting in a likelihood with all the usual and necessary regularity conditions.

2.2.5 Prior distributions for Bayesian meta-analyses

When using the Bayesian version of this methodology, the modelling must include a choice of prior distribution for μ, τ and ϕ. We chose vague priors that were constant when the parameter was transformed to lie on the whole of the real line, using a logarithmic transformation for positive quantities such as τ, ν, c or δ. This corresponds to the vague prior $g_\tau(\tau) \propto 1/\tau$, etc. Smith, Spiegelhalter et al. [11] discuss choice of Bayesian prior distributions for meta-analysis.

2.3 Numerical analysis and computing

Although practitioners might baulk at computing a likelihood function that contains an integral, a univariate integral is trivial in today's world of fast computers. A fortran95 program, available on request, was written by the authors to do the analysis. We also used the R language for this, and this is probably the best approach for practitioners. The program had to maximise the likelihood function, to compute the covariance matrix, and to find confidence intervals from the profile likelihood function. The program was also able to carry out a Bayesian analysis. The NAG library routines were used to evaluate integrals, maximise functions, invert matrices, and generate random numbers. A difficulty is that ℓ must be calculated accurately for any choice of parameter values that the function minimiser chooses, which requires robust computations. A particular problem arose with the t-distribution for very large numbers of degrees of freedom, where Fisher's expansion of the logarithm of the pdf was used [18]. Bayesian computations are nowadays often done by Markov chain Monte Carlo (MCMC). This has the nice feature of yielding a random sample from the posterior distribution of the parameter(s) of interest, from which means and credibility intervals can be read off. The real strength of MCMC is that one can sample from a multivariate pdf. However, in many applications this is not necessary, and there is the disadvantage that the Markov chain has to 'warm up'. MCMC is a powerful but computationally very expensive method. Hence we used importance sampling Monte Carlo (see e.g. Geweke [19]) rather than MCMC to evaluate the posterior distribution in Eq. 3. This has the advantage of not requiring a 'burn in' period. Here, the sampling function h was the multivariate-normal approximation to the product of the likelihood and the prior pdfs, derived in the frequentist analysis. This is available, as the likelihood maximum gives the distribution mean, and the Hessian can be inverted to give the covariance matrix. Note that in the space of parameters transformed to lie on the whole real line, the prior distributions are constant, although this is not necessary for our approach to work. Random vectors from the multivariate normal pdf h were generated, which is readily done by a Cholesky decomposition of the covariance matrix $\mathbf{V}$. If $\mathbf{V} = \mathbf{L}\mathbf{L}^T$, where $\mathbf{L}$ is lower triangular, then an i.i.d. unit normal random vector $\mathbf{X}$ is converted to $\mathbf{Y} = \mathbf{L}\mathbf{X}$, where $\mathbf{Y}$ has covariance matrix $\mathbf{V}$.

Each random vector $\mathbf{Y}$ is given a weight of $w(\mathbf{Y}) = \mathcal{L}(\mathbf{Y})/h(\mathbf{Y})$, so that a good approximation to the likelihood function will yield roughly equal weights for each random vector. Since flat priors have been adopted the weighted values of $\mathbf{Y}$ approximate the shape of the joint posterior distribution for the three model parameters. The posterior mean and the credibility interval for μ were then computed from the normalised weighted distribution of the simulated $\mathbf{Y}$, the latter being obtained as the 2.5 and 97.5% points of the univariate posterior distribution. Simulated $\mathbf{Y}$ and the corresponding $w(\mathbf{Y})$ had to be sorted, and the quickersort algorithm was used. After 20,000 simulations, means and credibility intervals had stabilised. This is probably a lazy way to evaluate a posterior distribution, but has the merit of being easy to program. In particular, the computation of the credibility region becomes very simple. An attempt was made to make the method adaptive, by using the simulations to refine the estimate of the 'best' covariance matrix $\mathbf{V}$, but this produced little improvement.

It must be pointed out that a poor choice of the likelihood maximum and of the covariance matrix does not invalidate the results, but it can drastically increase

Table 1 Estimated effect and confidence interval (credibility interval for Bayesian method), scale factor, appropriate shape parameter, and chi-squared for improvement in model fit for the CDP data set

Distribution	$\hat{\mu}$	CI	$\hat{\tau}$	3rd param	$\chi^2[1]$
Normal	0.389	0.073, 0.766	0.383	n/a	n/a
t	0.195	0.053, 0.361	0.0067	$\nu^{-1} = 2.02$	8.28
$\sinh^{-1}$	0.195	0.053, 0.336	0.0058	$c = 632$	8.48
Subbotin	0.199	0.0474, 0.371	0.33×10^{-6}	$\delta = 10.2$	7.29
t (Bayes)	0.188	0.175, 0.353	$.21 \times 10^{-3}$	$\nu^{-1} = 0.26$	n/a

Estimates are means of posterior distributions for the Bayesian analysis.

the computing time needed to obtain them, because the weights become erratic.

3 Results

Although many data sets have been examined, results are presented for only three, one small but typical data set with essentially just one outlier, a large data set with several outliers, and a large data set with no obvious outliers.

3.1 CDP-choline for cognitive and behavioural disturbances

Fioravanti and Yanagi [20] present a meta-analysis of cytidinediphosphocholine (CDP-choline) for cognitive and behavioural disturbances associated with chronic cerebral disorders in the elderly. There are ten studies for the outcome used here, memory measures, and the y_i are standardised mean differences of treatment versus placebo. Positive values of y_i indicate that the treatment is beneficial.

Table 1 shows the estimated mean effect and its confidence interval using the various models proposed for the random effect. Here 'normal' refers to the use of the conventional normal model for the random effect. The score statistic (5) was 2.28, which suggests that a long-tailed distribution is needed and hence results using the short tailed beta distribution are not shown in Table 1. Some illustrative results from a Bayesian meta-analysis are also shown in Table 1 using the t distribution for the random effect.

It is clear that the 8th study in the forest plot of Fig. 1 is an outlier. The three long-tailed distributions all appear to describe the data better than the normal random effects model, as shown by the chi-squared of

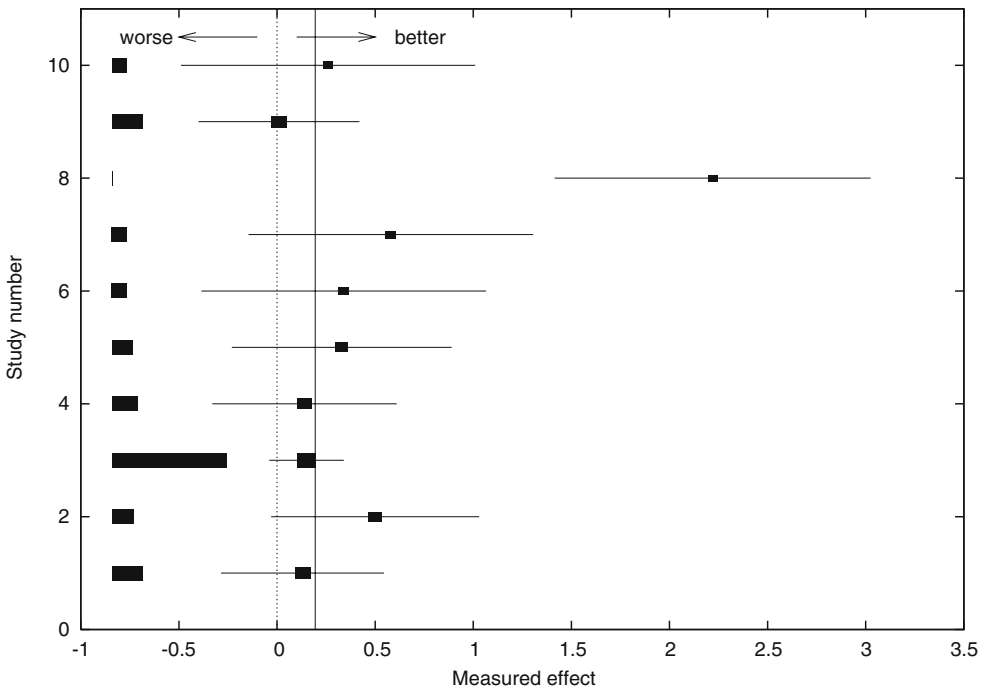

Fig. 1 Forest plot for the CDP study, showing ratios of weights for the t distribution and the normal distribution for the random effect

Table 2 Estimated effect and confidence interval (credibility interval for Bayesian method), scale factor, appropriate shape parameter, and chi-squared for improvement in model fit for the fluoride toothpaste data set

Distribution	$\hat{\mu}$	CI	$\hat{\tau}$	3rd param	$\chi^2[1]$
Normal	−0.300	−0.341, −0.262	0.119	n/a	n/a
t	−0.280	−0.313, −0.247	0.0487	$\nu^{-1} = 0.87$	28.92
$\sinh^{-1}$	−0.280	−0.3126, −0.247	0.045	$c = 42.6$	28.46
Subbotin	−0.279	−0.313, −0.248	0.48×10^{-4}	$\delta = 10.2$	7.00
Subbotin (Bayes)	−0.281	−0.319, −0.231	$.31 \times 10^{-5}$	$\delta = 5.61$	n/a

Estimates are means of posterior distributions for the Bayesian analysis.

about 8 on 1 degree of freedom. The forest plot shows how the influence of the outlier has been reduced. The rectangles to the left of the more conventional forest plot in Fig. 1 show the ratio of weights w_i/w_i^0, evaluated as described in Section 2.1 using the fitted t distribution for the random effect, and similar results are also shown in the funnel plots for the two further examples described below. These weight ratios aid interpretation, for example these indicate that the t distribution gives much more weight to study 3, and precious little weight to the outlying study 8, compared to a more conventional random effects meta-analysis.

The Bayesian analysis does not differ much from the frequentist, even for this small number of studies. Clearly, the estimate of the effect $\hat{\mu}$ has roughly halved when using the alternative formulations of the random effects model, while remaining statistically significant. Estimates of $\hat{\mu}$ from the three long tailed distributions are very similar.

3.2 Fluoride toothpaste for preventing dental caries

Marinho et al. [21] present a meta-analysis of fluoride toothpastes for preventing dental caries in children and adolescents. There are 70 studies, and the outcome is the difference between treatment and control of tooth areas with caries; negative values of y_i indicate that the treatment is beneficial. This is a large meta-analysis with obvious outliers, but where the treatment benefit is not in doubt, and there was no suggestion of publication bias in the original review. Table 2 shows the results where, as there is no evidence that short tails are required for the random effect, the beta distribution is again not fitted; illustrative results from a Bayesian

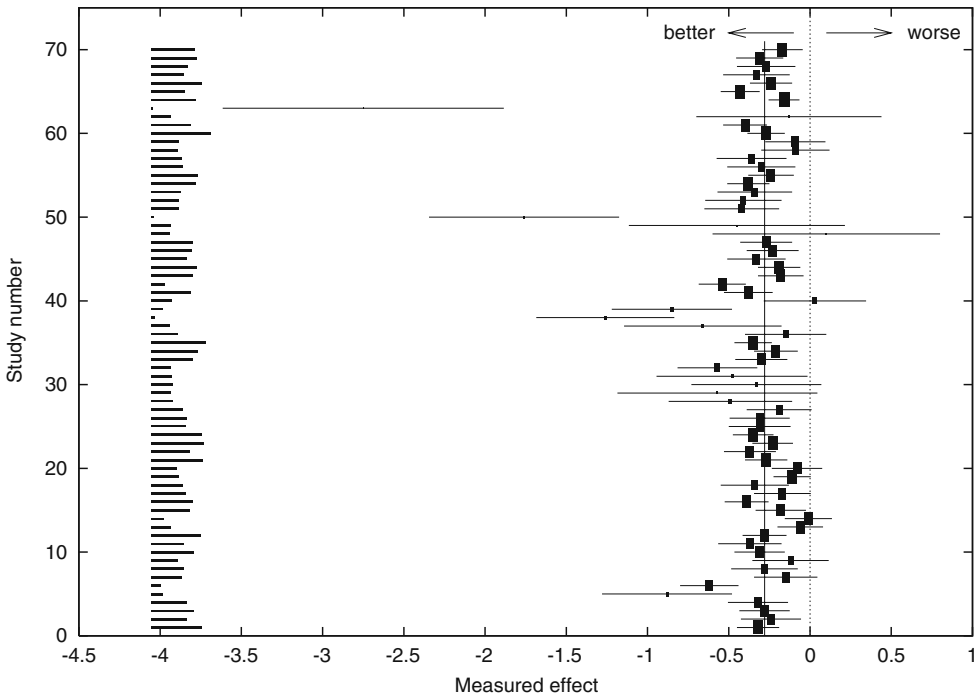

Fig. 2 Forest plot for the fluoride toothpaste study, showing ratios of weights for the arcsinh distribution and the normal distribution for the random effect

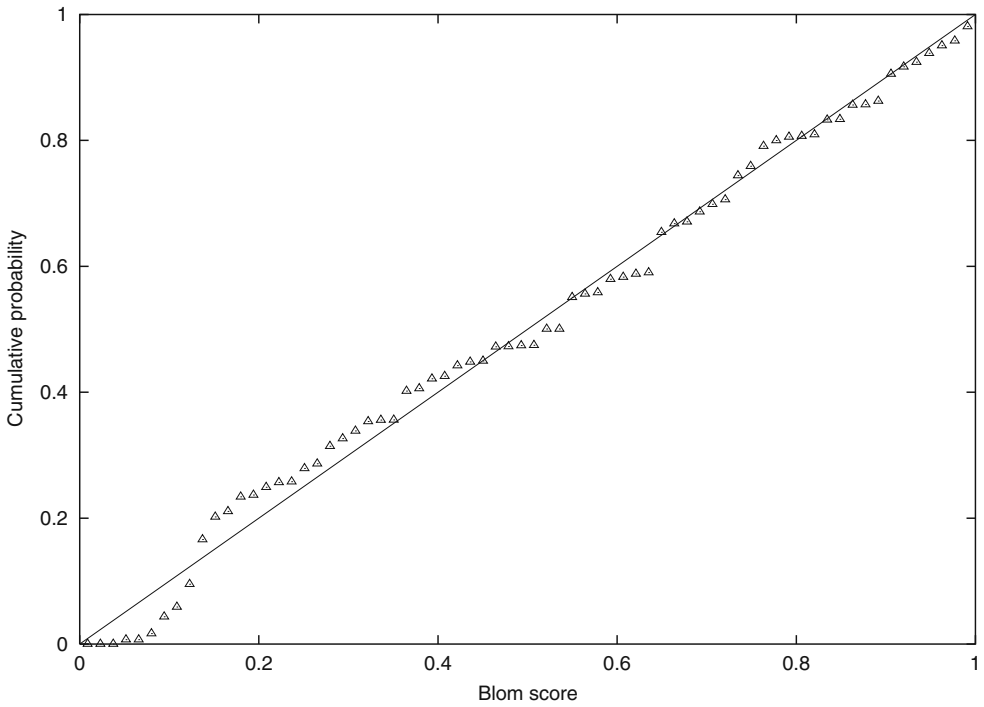

Fig. 3 Assessing goodness of model fit: distribution function from integrating Eq. 1 against expected value (Blom score) for the fluoride toothpaste data with the normal random effects model

meta-analysis are also shown in Table 1 using the Subbotin distribution.

The improvement in model fit using the long tailed distributions for the random effect is again considerable. The score statistic shows the correct sign of kurtosis, but is small at 1.33. The modified forest plot in Fig. 2 shows that the influence of several outliers has been reduced by using the arcsinh distribution for the random effect.

In order to further investigate the better model fit afforded by a random effects model with long tails, 'Blom' plots were produced for the conventional nor-

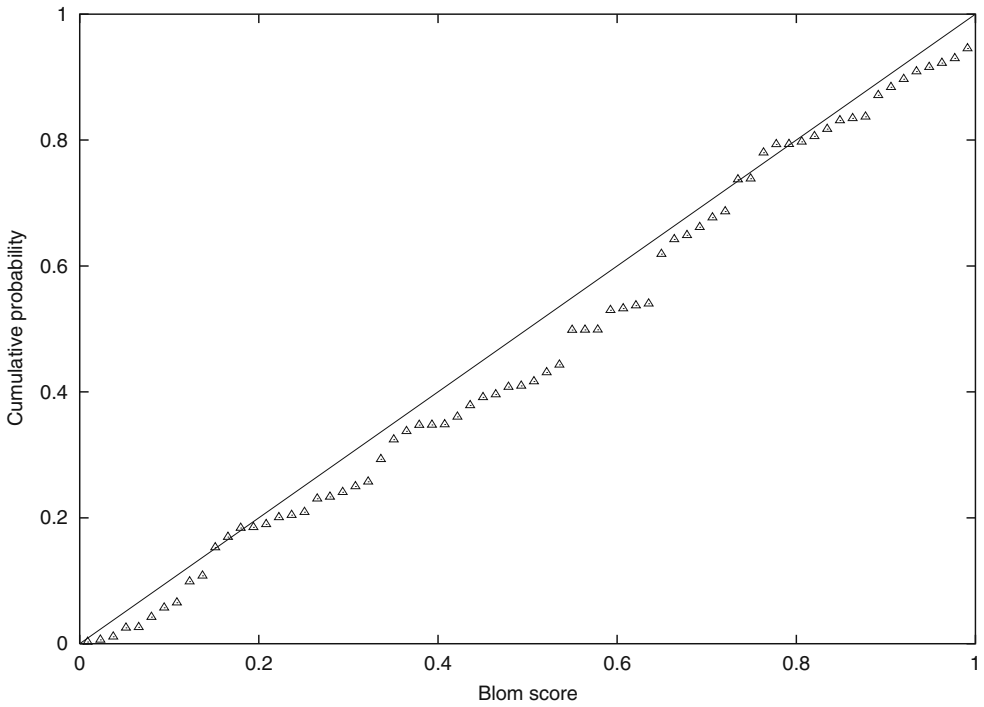

Fig. 4 Assessing goodness of model fit: distribution function from integrating Eq. 1 against expected value (Blom score) for the fluoride toothpaste data with the arcsinh random effects model

Table 3 Estimated effect and confidence interval (credibility interval for Bayesian method), scale factor, appropriate shape parameter, and chi-squared for improvement in model fit for the aspirin data set

Distribution	$\hat{\mu}$	CI	$\hat{\tau}$	3rd param	$\chi^2[1]$
Normal	1.247	1.087, 1.426	0.154	n/a	n/a
t	1.248	1.051, 1.446	0.158	$\nu^{-1} = 0.001$	0
$\sinh^{-1}$	1.248	1.052, 1.447	0.158	$c = 0.005$	0
Subbotin	1.249	1.065, 1.434	0.27	$\delta = 0.16$	0.05
Beta	1.251	1.064, 1.437	0.571	$\nu^{-1} = 10$	0.08
Normal (Bayes)	1.238	1.099, 1.409	$.16 \times 10^{-3}$	n/a	n/a

Estimates are means of posterior distributions for the Bayesian analysis.

mal random effects model and the arcsinh model in Figs. 3 and 4 respectively, which show a considerable improvement in model adequacy when using the arcsinh model, at the left of the plot. These Blom plots are a variation of the normal probability plots suggested by Hardy and Thompson [9], where the cumulative probability (the ordered values of the cumulative distribution functions, evaluated for the observed data and using the fitted model) is plotted against Blom scores, $(i - 3/8)/(n + 1/4)$ for $i = 1, 2, \cdots n$. If the model fits then the plotted points should approximately follow a straight line through the origin with gradient one, and this line is also shown in Figs. 3 and 4, in order to aid interpretation. Interestingly however, the model fit illustrated in the Blom plot in Fig. 4 appears slightly worse at the right. The distribution of results in fact seems to have outliers only in that some studies find very large benefit from fluoride toothpaste. This is a little puzzling. Maybe there is some publication bias, so that corresponding outliers to the right do not appear. It would be possible to model such effects with tailor-made skew distributions of η, but clearly one can not invent a new distribution for each analysis.

Here the change in the size of the estimated effect $\hat{\mu}$ when permitting the use of a random effects model with long tails is modest, essentially from -0.30 to -0.28. The use of long tailed distributions for the random effect do not bring the inference of an overall treatment effect into question.

3.3 Oral aspirin for acute pain

Edwards et al. [22] presented a meta-analysis of 63 studies of single dose oral aspirin for acute pain. The drug

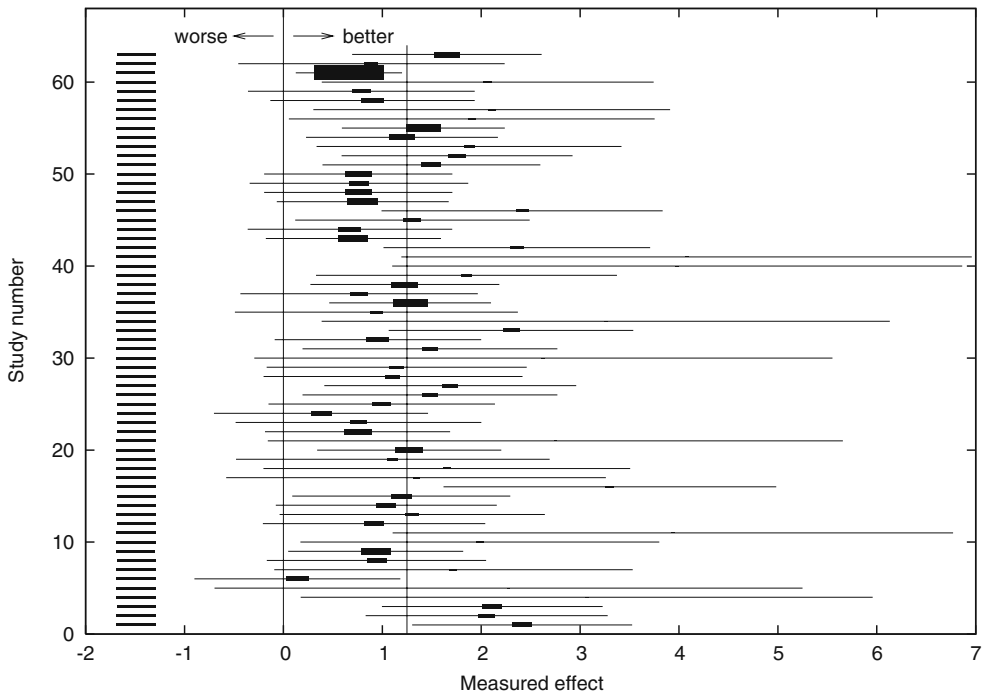

Fig. 5 Forest plot for the aspirin study, showing ratios of weights for the Subbotin distribution and the normal distribution for the random effect

was in 600 or 650 mg doses versus placebo, the outcome being patients with 50% or above pain relief. Values of $y_i > 0$ indicate that the treatment is beneficial.

Table 3 shows results for the aspirin data, where fitting long-tailed distributions does not change the results, and a short-tailed distribution was indicated by the score statistic of −0.97. Although the shape parameter δ of the Subbotin distribution indicates a fairly platykurtic distribution, there is very little drop in chi-squared, and this parameter is not well determined by the data. Whether one takes the minimum-AIC model as the usual random effects model, or adopts the 3-parameter model with the Subbotin distribution for η makes no difference to the estimated value $\hat{\mu}$, and its confidence interval is only very slightly widened on adopting the Subbotin model. Figure 5 shows the modified forest plot, for the Subbotin model, again confirming that it makes little change to the weights. Since there is some evidence that a distribution with short tails is required for the random effect, results are shown using the beta distribution in Table 3, where illustrative results from a Bayesian meta-analysis are also shown using the conventional normal distribution for the random effect. All of the various models provide strong evidence of a treatment effect, indicating that aspirin is much more effective for pain relief than a placebo.

4 Conclusions

The use of the proposed model, with a third parameter that allows the random effect to be long or short-tailed, has been shown to be able to reduce the weight of any outlying studies in a meta-analysis data set.

It is thus possible to allow the data to give the 'best' weight to an outlier, rather than arbitrarily removing it. When the distribution of the random effect is long-tailed, the conventional random effects model will fit with a high value of the random effect variance τ^2. As the weight of a study is $\frac{1}{\sigma_i^2+\hat{\tau}^2}$, large studies (with small σ_i^2) are not able to contribute as much as they should to the final consensus $\hat{\mu}$.

It becomes clear from examining large collections of studies that there may be several outliers, so that modelling their presence is desirable. The implications of these unusual study findings, for the estimated treatment effect and its confidence interval, can be large. The three long-tailed distributions used all gave similar results, and using more than one allows a sensitivity analysis to be done.

We have not found any studies that suggest strongly platykurtic distributions of random effect, so future modelling should focus on long-tailed distributions. The cholesterol-lowering data set described in Sutton et al. [3] gave a slightly platykurtic distribution, with a chi-squared reduction of 1.3 on using the Subbotin distribution for the random effect. The movement of the estimate $\hat{\mu}$ was very slight, i.e. from −0.119 to −0.123. Simulation showed that the log odds ratio, with the usual correction described in [3], differs from normality in being slightly skewed and having a slightly negative kurtosis. A small platykurty in the random effect may be a reflection of this.

Also, rounding of study results to few decimal places can introduce apparent platykurty, because a uniformly distributed rounding error is added to the normal distribution of results, making the distribution of y_i short tailed. It is of course important to quote results to only a few decimal places, but it is best to be cautious when transcribing such published data for analysis.

It is interesting, finally, to go back to the question of publication bias, another problem for the support of EBM by meta-analysis. We simulated two data sets of 100 'studies' with standard errors $\sigma_i = 1$, and a normal random effects distribution with $\tau = 1$. In one data set however studies with $y_i < -1$ were discarded, in order to introduce a simple publication bias mechanism. This data set fitted a long-tailed distribution, and in fact the kurtosis of a larger simulated sample was small but positive, at 0.13. This shows that publication bias can lead to long-tailed distributions of the random effect being fitted. It is not however possible to deduce publication bias from the existence of a long-tailed distribution for the random effect and the possibility of simultaneously modelling publication bias and long tailed distributions for the random effect is currently being investigated.

References

1. Copas J, Jackson D (2004) A bound for publication bias based on the fraction of unpublished studies. Biometrics 60:146–153
2. Baker RD, Jackson D (2006) Using journal impact factors to correct for the publication bias of medical studies. Biometrics 62:785–792
3. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F (2000) Methods for meta-analysis in medical research. Wiley, Chichester
4. Biggerstaff BJ, Tweedie RL (1997) Incorporating variability of estimates of heterogeneity in the random effects model in meta-analysis. Stat Med 16:753–768
5. Simonian R, Laird N (1986) Meta-analysis in clinical trials. Control Clin Trials 7:177–188
6. Hardy RJ, Thompson SG (1996) A likelihood approach to meta-analysis with random effects. Stat Med 15:619–629

7. Barnett V (1978) The study of outliers: purpose and model. Appl Stat 27:242–250
8. Fleiss JL (1993) The statistical basis of meta analysis. Stat Methods Med Res 2:121–145
9. Hardy RJ, Thompson SG (1998) Detecting and describing heterogeneity in meta-analysis. Stat Med 17:841–856
10. Lee KJ, Thompson SG (2008) Flexible parametric models for random effects distributions. Stat Med 27:418–434. Available online at http://www3.interscience.wiley.com/cgi-bin/abstract/114240283/ABSTRACT
11. Smith TC, Spiegelhalter DJ, Thomas A (1995) Bayesian approaches to random-effects meta-analysis: a comparative study. Stat Med 14:2685–2699
12. Van Houwelingen HC, Zwinderman KH, Stijnen T (1993) A bivariate approach to meta-analysis. Stat Med 12:2273–2284
13. Aitkin M (1999) Meta-analysis by random effect modelling in generalized linear models. Stat Med 18:2343–2351
14. Johnson NL, Kotz S, Balakrishnan N (1994) Continuous univariate distributions. Wiley, New York
15. Burnham KP, Anderson DR (1998) Model Selection and Inference: a practical information-theoretic approach. Springer, New York
16. Tarone RE (1985) Score tests. In: Kotz S, Johnson NL (eds) Encyclopaedia of Statistical Sciences. Wiley, New York.
17. Subbotin MT (1923) On the law of frequency of errors. Mathematicheskii Sbornik 31:296–301
18. Fisher RA (1925) Expansion of "Student's" integral in powers of n^{-1}. Metron 5:109–120
19. Geweke J (1991) Genereric, algorithmic approaches to Monte-Carlo integration in Bayesian inference. In: Flournoy N, Tsutakawa RK (eds) Statistical Multiple Integration, pp 117–135
20. Fioravanti M, Yanagi M (2005) Cytidinediphosphocholine (CDP-choline) for cognitive and behavioural disturbances associated with chronic cerebral disorders in the elderly (Review). The Cochrane Collaboration, www.cochrane.org/reviews/en/ab000269.html
21. Marinho VCC, Higgins JPT, Logan S, Sheiham A (2002) Fluoride toothpastes for preventing dental caries in children and adolescents. The Cochrane Collaboration, www.cochrane.org/reviews/en/ab002278.html
22. Edwards JE, Oldman A, Smith L, Collins SL, Carroll D, Wiffen PJ, McQuay HJ, Moore RA (2008) Single dose oral aspirin for acute pain (meta-analysis). The Cochrane Collaboration, www.cochrane.org/reviews/en/ab002067.html