Article:
Walwyn, R and Roberts, C (2017) Meta-analysis of standardised mean differences from randomised trials with treatment-related clustering associated with care providers. Statistics in Medicine, 36 (7). pp. 1043-1067. ISSN 0277-6715

https://doi.org/10.1002/sim.7186

Full Title:

Meta-analysis of standardised mean differences from randomised trials with treatment-related clustering associated with care providers

Short Title:

Meta-analysis of standardised mean differences from clustered trials

Authors:

Rebecca Walwyn (University of Leeds)

Chris Roberts (University of Manchester)

Contact Information for Corresponding Authors:

Rebecca Walwyn, Leeds Institute for Clinical Trials Research, University of Leeds, Leeds, United Kingdom, LS2 9JT.

Email: R.E.A.Walwyn@leeds.ac.uk

Keywords: standardised mean differences; meta-analysis; therapist effects

Acknowledgements: Rebecca Walwyn was funded by a Medical Research Council Special Training Fellowship in Health Services and Health of the Public (ref: G0501886). The authors would like to thank Pamela Gillies, Clair Chilvers, Michael Dewey, Karin Friedli, Ian Harvey, Adrian Hemmings, Michael King, Peter Bower, Roslyn Corney and Sharon Simpson for access to the datasets used in the example and to the reviewers for their helpful suggestions.

Abstract

In meta-analyses, where a continuous outcome is measured with different scales or standards, the summary statistic is the mean difference standardised to a common metric with a common variance. Where trial treatment is delivered by a person, nesting of patients within care providers leads to clustering that may interact with, or be limited to, one or more of the arms. Assuming a common standardising variance is less tenable and options for scaling the mean difference become numerous. Metrics suggested for cluster-randomised trials are within, between and total variances. For unequal variances, the control arm or pooled variances. We consider summary measures and individual-patient-data (IPD) methods for meta-analysing standardised mean differences (SMDs) from trials with two-level nested clustering, relaxing independence and common variance assumptions, allowing sample sizes to differ across arms. A general metric is proposed with comparable interpretation across designs. The relationship between the method of standardisation and choice of model is explored, allowing for bias in the estimator and imprecision in the standardising metric. A meta-analysis of trials of counselling in primary care motivated this work. Assuming equal clustering effects across trials, the proposed random-effects meta-analysis model gave a pooled SMD of -0.27 (95% CI -0.45 to -0.08) using summary measures and -0.26 (95% CI -0.45 to -0.09) with the IPD. While treatment-related clustering has rarely been taken into account in trials, it is now recommended that it is considered in trials and meta-analyses. This paper contributes to the uptake of this guidance.

1. INTRODUCTION

Summary measures approaches to statistical pooling or 'meta-analysis' of randomised trials first involve extracting a summary statistic, representing a treatment effect, from each trial and then calculating a weighted average of them [1, 2]. Where the outcome is normally-distributed, for example the severity of depression, the summary statistic is often an absolute mean difference. If the outcome is measured with different scales or standards across trials, for instance with the HADS-D [3], PHQ-9 [4] and the BDI [5], then the relevant summary statistic is the absolute mean difference, standardised to a common metric. That is, the standardised mean difference (SMD) or effect size. Outcomes are then assumed to be linearly equitable across trials, regardless of the measurement tool used, and the summary statistic is interpreted as a mean difference given in units of a standard deviation (SD) [6]. Where outcomes can be assumed to be independent and their SD homogeneous, the population SMD is defined as the difference in means across arms, divided by the common SD of the outcome.

Independence and common variance assumptions are less tenable when the treatment a patient receives is delivered by a health professional, such as in talking or physical therapies or surgery. Systematic variation or 'clustering' in patient outcomes by care provider arises when providers differ in characteristics related to outcome, such as training, skill, experience or empathy. As with cluster-randomised trials, the resulting correlation among outcomes within clusters violates the assumption of independence. However, treatment-related clustering also violates the common variance assumption. Provider characteristics may also differ across arms, for instance with greater skill or different training being required for one treatment than another. There may also be greater standardisation of one treatment, or one may be more established so that there is greater experience associated with it. The consequence of violations to the standard assumptions is that there is no longer a single common metric; the options available for scaling the mean difference being numerous. In general, each one is associated with a different population parameter and requires a different interpretation.

In cluster-randomised trials, typically, treatments are randomly allocated to entire clusters in a fully-nested, parallel-group design [7]. It is generally assumed that the clustering effect is homogeneous across treatment arms so the between- and within-cluster variances, which make up the total variance, are the same in both arms and a random intercept model appropriate for the analysis of each trial. Under this assumption, both White and Thomas [8] and Hedges [9] have suggested population SMDs based on the between, within and total SDs respectively. A between-cluster SD cannot be defined if there is only one cluster per arm in a trial. Similarly, where cluster-level analyses are reported, the within-cluster SD may not be. In both cases it would be possible to make assumptions about the intra-cluster correlation (ICC) and report and interpret the SMD in units of the total SD. While the choice of metric should depend on the inference of interest to a meta-analyst [9], SMDs based on the total and within SDs reduce to the standard SMD when outcomes are independent. If clustering is ignored in the published analyses, estimates of the between, within, and total SDs are unlikely to be readily available. Therefore, their population values may be difficult to estimate directly. To circumvent this problem, White and Thomas [8] and Hedges [9] suggest replacing the total SD by a 'naïve' SD, given by the total mean squares, in estimating the total SD SMD, and correcting for a bias that arises in doing so.

The simple situation, in which independence and normality assumptions hold but the variances differ across two treatment arms, is classically referred to as the Behrens-Fisher problem [10]. Glass [11] argued that between-trial heterogeneity in the treatment arms obscures interpretation when pooling trials in this situation and recommended the control arm SD be used as the metric of choice if the comparator is no treatment. It is arguable that this advantage is lost if control content also varies from trial to trial. As an alternative, Huynh [12] suggested pooling the SDs across arms, using the effect size proposed by Cohen [13, p.44]. In contrast to the standard SD, sample SDs in this metric estimate different population SDs. While it has been argued that the resulting distribution is rather contrived, and requires careful interpretation [14], this SMD has the advantage of reducing to the standard SMD when the outcome SDs are homogeneous across arms, utilising all available outcome

data, and minimises the small-sample bias in the trial SMDs identified by Hedges [6, 15]. Huynh [12] assumed the sample size is the same in each arm. Where it differs, as is often the case, we propose a more general pooled outcome SD that could be used, weighting the SDs by the sample size in each arm. A further option might be to use the associated baseline SD, a metric more commonly recommended for standardised mean change scores [16-20]. This may appeal particularly where eligibility criteria are similar across studies.

Use of individual-patient-data (IPD) in meta-analyses of SMDs appears to be limited, but see [21-23] for examples. Goldstein et al [24] described an IPD approach with the level-1 or within-cluster SD as the common metric. This was illustrated using studies of class size where students were nested within classes, schools and studies, and small versus large class size represented the treatment arms. The inclusion of a further level in the meta-analysis makes their approach especially relevant but, in contrast, schools are crossed with arms in their example. And, while they allude to models that allow for between-arm or trial heteroscedasticity, they do not consider nested study designs, the rationale or implications of the choice of metric, imprecision in the standardising SD, or the relationship between the method for standardising outcome data and the choice of model for the meta-analysis.

This paper proposes summary measures and IPD approaches to the meta-analysis of standardised mean differences from randomised trials with uniform two-level nested designs and treatment-related clustering. It builds on earlier work [25], addressing the simpler situation in which absolute mean differences are to be pooled, but due to the additional complexities here, between-trial homogeneity in the within-trial clustering effects is assumed throughout. In both papers, the nested designs considered assume there is a single therapist-per-patient. In a fully nested design different care providers deliver every treatment, while at least one treatment does not require care providers in a partially nested design (see Walwyn and Roberts [26] for further description of the full range of therapist designs). The statistical model recommended at the trial-level for both nested designs is a two-level heteroscedastic model [27]. This includes a random effect for the care provider but allows the provider and patient level variances

to differ across arms, constraining the provider variance to zero in arms with no care providers for partially nested designs.

We begin in section 2 by outlining the example that motivated this work. In section 3 we set out the summary measures approach proposed by Hedges [6, 15], highlighting the steps involved a standard meta-analysis of SMDs and how it differs from a meta-analysis of absolute mean differences. In section 4 we extend this approach, and those suggested by White and Thomas [8], Hedges [9] and Huynh [12], proposing a general metric that simultaneously relaxes independence and between-arm common variance assumptions, also allowing the number of patients to differ across arms. In section 5 we first outline the steps suggested by Goldstein et al [24] for a standard IPD meta-analysis of SMDs, highlighting how these could be modified to allow for imprecision in the standardising metric. We then extend them, initially relaxing the between-arm homoscedasticity assumption for the Behrens-Fisher case, and then simultaneously relaxing the independence and between-arm homoscedasticity assumptions necessary to pool trials with treatment-related clustering. In section 6 we illustrate our methods using our motivating example, concluding in section 7 with a discussion and limitations.

2. MOTIVATING EXAMPLE

As in our previous paper on the meta-analysis of absolute mean differences [25], we were motivated by Bower and Rowland's [28] systematic review of the clinical and cost-effectiveness of counselling in UK primary care, which included 8 trials. As it is usual for counsellors to apply eclectic therapeutic approaches to a very wide range of social and clinical problems, the implications of therapist variation [26] are especially pertinent in this setting. The largest meta-analysis involved 7 trials [29-35] comparing counselling plus care from a general practitioner (GP) to just GP care using short term outcomes measuring the level of mental health symptoms.

Four of these trials [30, 31, 34, 35] reported the Beck Depression Inventory (BDI) [5], allowing a meta-analysis of the absolute mean differences [25]; the other three trials

[29, 32, 33] reported the General Health Questionnaire (GHQ) [36], the depression subscale of the Hospital Anxiety and Depression Scale (HADS-D) [3] and a short Symptom Index, respectively. All are commonly used self-report questionnaires: the BDI and the HADS-D measure severity of depression while the GHQ and Symptom Index are global measures of wellbeing. The BDI has 21 items (total scores 0 to 63) and the HADS has 7 items relating to depression (subscale scores 0 to 21) with higher scores indicating greater severity. The GHQ had 28 items (total scores 0 to 28) with a score above 4 indicating the presence of distress. The Symptom Index had 18 items (mean scores 0 to 4) with a norm of 0.61. Across these scales, a change of 0.5 SDs is generally regarded minimally important.

To pool all 7 trials, it was necessary to first transform the data on all four scales to a common metric. The published meta-analysis [28] gave an SMD of -0.24 SDs (95% CI -0.38 to -0.10), so, according to Cohen's [13] classification, the pooled treatment effect can be regarded as clinically small but statistically significant. Authors of the Cochrane review concluded 'counselling is associated with modest improvement in short-term outcome' and that it 'may be a useful addition to mental health services in primary care' [37]. Ignoring the co-intervention of GP care, each trial can be viewed as having a partially nested design, with counsellors delivering treatment in the intervention but not in the control arm. Across trials, there was a single counsellor per patient. The published meta-analysis used a standard summary measures approach, assuming independence of patient outcomes within trials and a common variance across arms, fitting a fixed-effects meta-analysis model assuming a common underlying SMD across trials.

3. THE STANDARD SUMMARY MEASURES APPROACH

Any summary measures meta-analysis of SMDs, or absolute mean differences for that matter, requires the systematic reviewer to assume outcomes are normally-distributed and to extract the sample means ($\bar{y}_{kh}$), SDs (s_{kh}) and sizes (n_{kh}) for each arm ($k = 0,1$) of every trial ($h = 1,\ldots, H$). Assuming these are all available from published or unpublished trial reports or direct correspondence with authors, the first

step is to choose an appropriate metric or SD for scaling the absolute mean difference. The SMD is thus a ratio, where the absolute mean difference is not. In the standard but also simplest scenario, where independence and homoscedasticity is assumed within and across each trial to be pooled, there is only one option for such a metric, the common SD σ . The population SMD is therefore defined [6, 15] as

$$\theta_{\text{SMD}} = \frac{\mu_1 - \mu_0}{\sigma} \qquad (1)$$

with the difference in population means of the treatment and control arm respectively in the numerator and a common standardising metric in the denominator. The second step specific to meta-analyses of SMDs is to choose an estimator for the denominator. Usually the population metric can be estimated, within each trial, by the sample SD in the treatment or control arm, s_{1h}, s_{0h} or by the sample SD pooled across arms, $s_{\bullet h}$. A third step is then to determine the sampling distribution for the chosen estimator of the denominator of the SMD, obtaining the relevant degrees of freedom. This tends to be the pooled sample SD $s_{\bullet h}$ because it maximises the degrees of freedom available, utilising all of the available data. As in the standard case the pooled sample variance is simply the mean squares error, that is $s^2_{\bullet h} = \text{MSE}_h$, it follows that its sampling distribution is exactly proportional to a chi-square with $n_{1h} + n_{0h} - 2$ degrees of freedom df_h. Since $s_{\bullet h}$ is a direct estimator of σ in the standard case, a fourth step of calculating a bias, relating to the choice of estimator for the metric, is avoided. Once one has a sampling distribution for the standardising metric, a fifth step, which is common to all meta-analyses, is to obtain the sampling distribution of the sample estimate of the population parameter so as to select an unbiased summary statistic and determine its standard error. There are two sample estimates of the population SMD. In the standard case, the first is commonly referred to as Cohen's d [13], the second as Hedges' g [15]. Cohen's d is the large-sample estimate of the population SMD and is given by simply replacing the population parameters by their sample equivalents in each trial as follows,

$$\hat{\theta}_{\text{Cohen's d},h} = \frac{\bar{y}_{1h} - \bar{y}_{0h}}{s_{\bullet h}} \qquad (2)$$

This makes no allowance for imprecision in the standardising SD, instead assuming that all the trial sample sizes are large so all the df_h are also large. Its sampling distribution is given asymptotically [15] by

$$\hat{\theta}_{\text{Cohen's d,h}} \sim \mathrm{N}\left(\frac{\mu_{1\mathrm{h}}-\mu_{0\mathrm{h}}}{\sigma_\mathrm{h}}, \left(\frac{1}{\mathrm{n}_{1\mathrm{h}}}+\frac{1}{\mathrm{n}_{0\mathrm{h}}}\right)+\frac{\theta^2_{\mathrm{SMD}}}{2\left(\mathrm{n}_{1\mathrm{h}}+\mathrm{n}_{0\mathrm{h}}\right)}\right) \qquad (3)$$

As Cohen's d is a ratio of a mean difference to a standardising SD, its standard error is equal to the standard error of the absolute mean difference plus a term that relates to the SD, the latter depending on the population parameter. As such, in contrast to a meta-analysis of absolute mean differences, in a meta-analysis of SMDs, the sample estimate and its standard error can be seen to depend on the variance of the outcome. If any of the trial sample sizes ($\mathrm{n}_{\bullet\mathrm{h}}$) are small, and particularly if $\mathrm{df}_\mathrm{h} \leq 10$, Hedges [15] showed that Cohen's d is biased for θ_{SMD}, and derived an alternative estimator, to correct for this,

$$\hat{\theta}_{\text{Hedges' g,h}} = \mathrm{c}\left(\mathrm{df}_\mathrm{h}\right)\left(\frac{\bar{\mathrm{y}}_{1\mathrm{h}}-\bar{\mathrm{y}}_{0\mathrm{h}}}{\mathrm{s}_{\bullet\mathrm{h}}}\right) \sim \mathrm{N}\left(\frac{\mu_{1\mathrm{h}}-\mu_{0\mathrm{h}}}{\sigma_\mathrm{h}}, \frac{\mathrm{c}\left(\mathrm{df}_\mathrm{h}\right)^2 \mathrm{df}_\mathrm{h}\left(1+\tilde{\mathrm{n}}_\mathrm{h}\theta^2_{\mathrm{SMD}}\right)}{\left(\mathrm{df}_\mathrm{h}-2\right)\tilde{\mathrm{n}}_\mathrm{h}}-\theta^2_{\mathrm{SMD}}\right) \qquad (4)$$

where $\tilde{\mathrm{n}}_\mathrm{h} = \left(\frac{1}{\mathrm{n}_{1\mathrm{h}}}+\frac{1}{\mathrm{n}_{0\mathrm{h}}}\right)^{-1}$ and $\mathrm{c}\left(\mathrm{df}_\mathrm{h}\right) = \Gamma\left(\frac{\mathrm{df}_\mathrm{h}}{2}\right) \Big/ \left(\sqrt{\frac{\mathrm{df}_\mathrm{h}}{2}}\,\Gamma\left(\frac{\mathrm{df}_\mathrm{h}-1}{2}\right)\right) \approx 1-\frac{3}{4\mathrm{df}_\mathrm{h}-1}$

Hedges' g is the unbiased estimate of the population SMD, appropriate regardless of the degrees of freedom, df_h, available for estimating the standardising metric, $\mathrm{s}^2_{\bullet\mathrm{h}}$. It is therefore preferred over Cohen's d. Hedges' g converges to Cohen's d as the trial sample sizes increase but is uniformly smaller than Cohen's d otherwise [15]. Since Hedges [6, 15] originally suggested substituting $\hat{\theta}^2_{\text{Hedges' g,h}}$ for θ^2_{SMD} when estimating the standard error, this is widely done in software. White and Thomas [8, p.150] showed that this introduces bias because the expectation of a squared estimate is equal to the squared parameter plus the variance of the estimate, not simply the squared

parameter (that is $\mathrm{E}\left(\hat{\theta}^2\right)=\theta^2+\sigma^2_{\hat{\theta}}$). So they proposed a refined estimator for the exact sampling variance, given by

$$\hat{\sigma}^2_{\hat{\theta}_{\text{Hedges' g,h}}}=\left(\frac{1}{\mathrm{n}_{\mathrm{h1}}}+\frac{1}{\mathrm{n}_{\mathrm{h0}}}\right)+\hat{\theta}^2_{\text{Hedges' g,h}}\left(1-\frac{\mathrm{df}_{\mathrm{h}}-2}{\mathrm{c}\left(\mathrm{df}_{\mathrm{h}}\right)^2\mathrm{df}_{\mathrm{h}}}\right) \qquad (5)$$

The first term clearly relates to the variance of the numerator of the trial SMD and the second to the variance of its denominator. This was originally derived by Hedges [38 p.391].

Again common to all meta-analyses, once one has an unbiased summary statistic (e.g. Equation (4)) and an unbiased estimate of its standard error (e.g. Equation (5)) a sixth step is to pool summary statistics using fixed or random-effects meta-analysis models [39]. The choice between a fixed or random-effects meta-analysis is based on whether it is reasonable to assume there is an underlying SMD common across trials. If it is, a fixed effects model may be fitted. In the more likely scenario where population SMDs vary across trials, a random-effects model should be fitted. Where there is substantial heterogeneity in SMDs across trials, it is important to explore possible explanations for this, perhaps in the context of a meta-regression [40], or give a tolerance interval for the effect in a new study [41]. Methods to allow for between-trial heterogeneity are beyond the scope of this paper, however. The uniformly minimum variance unbiased estimate (UMVUE) of any pooled treatment effect θ is given by the following weighted average [42, 43]

$$\hat{\theta}=\frac{\sum_{\mathrm{h}=1}^{\mathrm{H}}\mathrm{w}_{\mathrm{h}}\hat{\theta}_{\mathrm{h}}}{\sum_{\mathrm{h}=1}^{\mathrm{H}}\mathrm{w}_{\mathrm{h}}} \qquad (6)$$

where the trial weights, w_{h}, are the inverse of the sampling variance of the summary statistic. In a random-effects meta-analysis model, the weights are the inverse of the total sampling variance, given by the sum of the within, $\hat{\sigma}^2_{\hat{\theta}_{\text{Hedges' g,h}}}$, and between, $\hat{\tau}^2_{\mathrm{D-L}}$, trial variances, where the latter is often estimated using DerSimonian-Laird's (D-L) [44] methods of moments estimator. In a fixed-effects meta-analysis the between trial

variance is simply zero. It is usual for $\sigma^2_{\hat{\theta}_h}$ and $\tau^2_{\theta_h}$ to be replaced by their respective estimators $\hat{\sigma}^2_{\hat{\theta}_h}$ and $\hat{\tau}^2_{\theta_h}$ here, although Sidik and Jonkman [45] suggest an alternative approach that is robust to sampling errors in the estimated weights.

A seventh step, recommended for summary measures SMD meta-analyses [6] but unnecessary when the summary statistic is an absolute mean difference, is to plug the pooled estimate of the population SMD from Equation (6) back into the estimate of the within trial variance given in Equation (5), which is used in estimating the trial weights, w_h, and to continue iterating until convergence. While this step is not always implemented in software, and it may not be desirable where there is any suspicion of heterogeneity in the SMD across trials, it is important because the initial weights depend on the trial estimate of the SMD. As the size of the weights increase as a function of SMD, particularly if the degrees of freedom available for estimating the standardising SD are low, omitting this step may lead to the pooled treatment effect being unduly affected by a single trial with an extreme SMD. Hedges [6] argued that when the degrees of freedom relating to the standardising SD are all large this step can be ignored.

Once you have converged estimates of the trial weights w_h, the eighth step is common to all meta-analyses. It is to calculate the standard error of the estimated parameter $\hat{\theta}$, simply by

$$\sigma_{\hat{\theta}} = \left(\sum_{h=1}^{H} w_h \right)^{-1/2} \qquad (7)$$

so that an approximate two-sided $100(1-\alpha)\%$ confidence interval for $\hat{\theta}$ is given by

$$\hat{\theta} \pm z_{1-\alpha/2}\sigma_{\hat{\theta}} \qquad (8)$$

4. A MORE GENERAL SUMMARY MEASURES APPROACH

The eight steps outlined above, hold specifically for the situation in which independence, normality and homoscedasticity can be assumed. When the

assumptions of independence and homoscedasticity no longer hold, the first five steps need to be modified. That is, (i) an appropriate SD for scaling the absolute mean difference must be chosen, (ii) an estimator for this SD found, (iii) the sampling distribution for this SD estimate determined, in order to obtain the relevant degrees of freedom, (iv) a bias relating to the choice of SD estimator calculated, and (v) the sampling distribution of the SMD estimate determined in order to obtain an unbiased summary statistic and its exact standard error. The final three steps (six to eight) are the same as in any meta-analysis of SMDs.

4.1 Choice of SD for Scaling the Absolute Mean Difference

A general population SMD can be defined as

$$\theta_{\text{SMD}} = \frac{\mu_1 - \mu_0}{\sigma_{\text{den}}} \qquad (9)$$

with the form of σ_{den} depending on the choice of standardising SD. As we have seen, in relation to Cohen's d and Hedges' g, $\sigma_{\text{den}} = \sigma$ for the standard SMD. For Glass'[2] SMD, the standardising metric is the control arm SD ($\sigma_{\text{den}} = \sigma_0$, where $\sigma_0 \neq \sigma$). For Huynh's [12] SMD, the standardising metric is a simple average of the treatment and control arm SDs ($\sigma_{\text{den}} = \sqrt{(\sigma_1^2 + \sigma_0^2)/2}$). And for White and Thomas' [8] or Hedges' [9] SMDs, the standardising metrics are the total, within or between cluster SDs respectively ($\sigma_{\text{den}} = \sigma_{\text{T}}, \sigma_{\text{W}}$ or σ_{B}). The issue faced when pooling SMDs from trials with treatment-related clustering is the presence of between-arm heteroscedasticity. Here, the need for a common metric across arms results in two options. The first is a metric that has a direct interpretation. One option is Glass' SMD, another is a SMD based on the baseline SD. Here, the standardising metric has a clear interpretation but only where trial designs are comparable. The second is a metric that requires pooling potentially different population SDs. This could be the within, between or total SDs. We argue that the most general standardising SD is what we will refer to as the pooled total SD. The population SMD based on this is given by

$$\theta_{\text{ptotal}} = \frac{\mu_1 - \mu_0}{\sqrt{\frac{(n_1 - 1)\sigma_{T1}^2 + (n_0 - 1)\sigma_{T0}^2}{n_1 + n_0 - 2}}} \qquad (10)$$

It applies to fully-nested trials with treatment-related clustering but reduces to a range of other SMDs in more restrictive scenarios. For partially-nested trials, the total SD in the control arm is the standard control arm SD so $\sigma_{T0} = \sigma_0$ in Equation (10). If the sample size is unequal across arms in the Behrens-Fisher case $\sigma_{T0} = \sigma_0$ and $\sigma_{T1} = \sigma_1$ in Equation (10). If the sample size is equal in this case then Equation (10) reduces to Huynh's [12] SMD. If the sample sizes, within and between cluster SDs are all equal across the arms, then Equation (10) reduces to White and Thomas'[8] and Hedges' [9] SMD based on σ_T. Consequently, all these metrics can be viewed as special cases of the more general metric proposed here, making the pooled total SMD an appealing option, compared to a pooled 'within' or pooled 'between' SMD. We will return to this issue in the discussion.

4.2 Choice of Estimator for the Standardising SD

White and Thomas [8] and Hedges [9] both give two options for estimating the total SD. Either it is estimated directly using the total SD (s_{Th}) in every trial ($h = 1, \ldots, H$), or if clustering is ignored in published analyses, and estimates of the total SDs are not readily available, it is estimated indirectly using the naïve SD ($s_{\bullet h}$), given, as with the standard case (Equations (2) and (4)), by the total mean squares. Sample estimates of the pooled total SD for the fully nested case are therefore given by

$$s_{T\bullet h} = \sqrt{\frac{(n_{1h} - 1)s_{T1h}^2 + (n_{0h} - 1)s_{T0h}^2}{n_{1h} + n_{0h} - 2}} \qquad (11)$$

or

$$s_{\bullet h} = \text{MST}_{\bullet h} = \sqrt{\frac{\text{SSW}_{1h} + \text{SSB}_{1h} + \text{SSW}_{0h} + \text{SSB}_{0h}}{n_{1h} + n_{0h} - 2}} \qquad (12)$$

respectively, where SSW_{kh} and SSB_{kh} are sums of squares within- and between-clusters in each arm ($k = 0,1$) of every trial ($h = 1,\ldots, H$), assuming cluster sizes are equal within trial arms. The sample SD in both Equations (11) and (12) is now a linear combination of mean squares terms rather than simply a single mean squares term (i.e. the mean square error) as it was assuming independence and homoscedasticity. Estimators of the pooled total and naïve SD are given in Table I, in terms of the sums of squares (SS), for the fully nested case and under the more restrictive scenarios. Note that the term for the total mean squares is biased for the pooled total variance, where clustering is present (see Section 4.4 for details of the implications of this).

[Insert Table I about here]

4.3 Sampling Distributions for Estimators of the Standardising SD

One consequence of the sample SDs in Equations (11) and (12) now being linear combinations of mean squares terms is that their sampling distributions are also no longer exactly proportional to chi-squares with $n_{1h} + n_{0h} - 2$ degrees of freedom df_h. Instead, they have sampling distributions approximately proportional to chi-squares with degrees of freedom given, using a Satterthwaite approximation [46] by

$$df_{s^2_{T\bullet h}} \approx \frac{\left((n_{1h}-1)(\hat{\rho}_{0h}-1)s^2_{W1h} + (n_{0h}-1)(\hat{\rho}_{1h}-1)s^2_{W0h}\right)^2}{(\hat{\rho}_{1h}-1)^2(\hat{\rho}_{0h}-1)^2\left(\sum_{k=0}^{1}\left(\frac{(n_{kh}-1)^2}{m_{kh}^2}\left(\frac{m_{kh}-1}{C_{kh}} + \frac{(1+(m_{kh}-1)\hat{\rho}_{kh})^2}{(C_{kh}-1)(\hat{\rho}_{kh}-1)^2}\right)s^4_{Wkh}\right)\right)} \quad (13)$$

and

$$df_{s^2_{\bullet h}} \approx \frac{\left(\sum_{k=0}^{1}(n_{kh}-1)s^2_{kh}\right)^2}{\sum_{k=0}^{1}\left(\frac{(n_{kh}-1)^2 s^4_{kh}\left(n_{kh}\left(1+(m_{kh}-1)\hat{\rho}^2_{kh}\right)-\left(1+(m_{kh}-1)\hat{\rho}_{kh}\right)^2\right)}{\left((n_{kh}-1)-(m_{kh}-1)\hat{\rho}_{kh}\right)^2}\right)} \quad (14)$$

respectively, where $\hat{\rho}_{0h}$ and $\hat{\rho}_{1h}$ are estimated control and treatment ICCs respectively, s^2_{Wkh} estimated within-cluster variances, m_{kh} the cluster size, C_{kh} the number of clusters

and s^2_{kh} the naïve variances in arm $k = 0,1$ of trial $h = 1,\ldots, H$. Derivations are given in Appendix A and B respectively as supporting web materials.

It is possible to rewrite Equation (13) in terms of total variances by replacing $s^2_{Wkh} = \sigma^2_{Tkh}(1-\rho_{kh})$. The degrees of freedom given in Equations (13) and (14) simplify under more restrictive scenarios to those summarised in Table II. Under independence the degrees of freedom are equal for the pooled total and naïve variances. Huynh [12 p.21] gave the degrees of freedom for his pooled SD. We give the degrees of freedom for the more general Behrens-Fisher situation but also those valid under a random-intercept model for the pooled total SD, correcting typographical errors in White and Thomas [8 p.151] and Hedges [9 p.364], and for the pooled naïve SD as given by Hedges [9 p.156], thereby correcting a further typographical error in White and Thomas [8 p.151].

[Insert Table II about here]

4.4 Bias Relating to the Choice of SD Estimator

The expectation of the pooled naïve variance under a two-level heteroscedastic model is given by

$$E\left[s^2_{\bullet h}\right] = b_h \sigma^2_{T\bullet} = \left(1 - \frac{(m_{1h}-1)\rho_1\sigma^2_{T1} + (m_{0h}-1)\rho_0\sigma^2_{T0}}{(n_{1h}-1)\sigma^2_{T1} + (n_{0h}-1)\sigma^2_{T0}}\right)\sigma^2_{T\bullet} \qquad (15)$$

with $\rho_{0h} = 0$ and $\sigma^2_{T0h} = \sigma^2_{0h}$ where trials are partially-nested (see Appendix C under supporting web materials for the derivation). Under independence, where $\rho_1 = \rho_0 = 0$, it can be seen that $b_h = 1$ and the naïve SD is unbiased. Hedges [9] gave the bias under a random intercept model as

$$b_h = \left(1 - \frac{(m_h - 1)\rho_h}{n_h - 1}\right) \quad (16)$$

As before, the bias in Equation (16) is a special case of that given in Equation (15). In all cases, the naïve variance underestimates the total variance by a factor linked to the design effect.

4.5 Sampling Distributions of the SMD Estimates

Huynh [12 p.4-6] and Hedges [9 p.360-2] have derived sampling distributions for large-sample (extending Cohen's d) and unbiased (extending Hedges' g) estimators of SMDs based on pooled and total SDs, respectively. These have a similar form and can be extended to give yet more general sampling distributions for the pooled total SMD. Suppose that, for each of h trials, the absolute mean difference in outcome observed between the treatment and control arms is distributed as

$$\bar{y}_{1h} - \bar{y}_{0h} \sim N\left(\mu_{1h} - \mu_{0h}, \frac{\left[1+(m_{1h}-1)\rho_1\right]\sigma_1^2}{n_{1h}} + \frac{\left[1+(m_{0h}-1)\rho_0\right]\sigma_0^2}{n_{0h}} \right) \qquad (17)$$

where the expectation ($E[\bar{y}_{1h} - \bar{y}_{0h}]$) and sampling variance ($\sigma^2_{\bar{y}_{1h}-\bar{y}_{0h}}$) of the absolute mean difference are unknown but the $a_{Tkh} = 1+(m_{kh}-1)\rho_{kh}$ denote known constants. If $\sigma^2_{ptotal} = E[s^2_{den,h}]/b_h$, where $s_{den,h}$ is given by Equation (11) or (12) and b_h , given in Equation (15, also a known constant), the large-sample estimator of θ_{ptotal} is given by

$$\hat{\theta}_{LS,ptotal,h} = \frac{(\bar{y}_{1h} - \bar{y}_{0h})\sqrt{b_h}}{s_{den,h}} \sim t_{df_h,\varphi_h} \sqrt{\frac{s^2_{\bar{y}_{1h}-\bar{y}_{0h}}}{s^2_{den,h}}} \qquad (18)$$

where t_{df_h,φ_h} is a non-central t-distribution with degrees of freedom df_h given in Table II and non-centrality parameter equal to $\varphi_h = \hat{\theta}_{ptotal} \Big/ \sqrt{\left(s^2_{\bar{y}_{1h}-\bar{y}_{0h}} / s^2_{den,h}\right)}$. A derivation for Equation (18) is given in Appendix D as supporting web materials.

Again Equation (18) simplifies. In the standard case, Equation (18) is simply Cohen's d ($\rho_1 = \rho_0 = 0, a_{Tkh} = 1,\ b_h = 1,\ \sigma_1^2 = \sigma_0^2,\ s^2_{\bar{y}_{1h}-\bar{y}_{0h}} / s^2_{den,h} = 1/n_{1h} + 1/n_{0h}$). It is Huynh's [12 p.4] g where $a_{Tkh} = 1,\ b_h = 1$, and $s^2_{\bar{y}_{1h}-\bar{y}_{0h}} / s^2_{den,h}$ is equal to Huynh's k^2. It is White and Thomas' [8 p.150] g_{un} if $b_h = 1, \sigma^2_{B1} = \sigma^2_{B0}, \sigma^2_{W1} = \sigma^2_{W0}$ and $s^2_{\bar{y}_{1h}-\bar{y}_{0h}} / s^2_{den,h} = Var[G]$, i.e. $deff_h(1/n_{h1} + 1/n_{h0})$ or $deff_{h1}/n_{h1} + deff_{h0}/n_{h0}$. Finally, Equation (18) is Hedges' [9 p.360] D where $a = a_{Tkh}, \sigma^2_{B1} = \sigma^2_{B0}, \sigma^2_{W1} = \sigma^2_{W0}$ and $s^2_{\bar{y}_{1h}-\bar{y}_{0h}} / s^2_{den,h}$ is equal to his $a/\tilde{N}$.

It follows from the definition of a non-central t-distribution (see Huynh [12 p.4] and White and Thomas [8 p.150]) that, where $df_h > 2$,

$$E\left[\hat{\theta}_{LS,h}\right] = \theta_{SMD}/c(df_h) \text{ and}$$

$$\sigma^2_{\hat{\theta}_{LS,h}} = \left(\frac{df_h}{df_h - 2}\right)\left(\frac{s^2_{\bar{y}_{1h}-\bar{y}_{0h}}}{s^2_{den,h}} + \theta^2_{SMD}\right) - E\left[\hat{\theta}_{LS,h}\right]^2 \quad (19)$$

with the asymptotic standard error of $\hat{\theta}_{LS,h}$ given by

$$\hat{\sigma}_{\hat{\theta}_{LS,h}} = \sqrt{\frac{s^2_{\bar{y}_{1h}-\bar{y}_{0h}}}{s^2_{den,h}} + \frac{\hat{\theta}^2_{SMD}}{2df_h}} = \sqrt{\frac{s^2_{\bar{y}_{1h}-\bar{y}_{0h}}}{s^2_{den,h}} + \frac{c_h\hat{\theta}^2_{SMD}}{2b^2_h}} \quad (20)$$

correcting typographical errors in Huynh [12 p.5] and Hedges [9 p.361]. Note that, in Hedges [9], $df_h = b^2_h/c_h$ due to his use of Box's [47] generalisation of Satterthwaite's approximation [46] for the degrees of freedom. The result in Equation (19) implies that the unbiased estimator of θ_{ptotal} is

$$\hat{\theta}_{ptotal,h} = c\left(df_{s^2_{den,h}}\right)\left(\frac{(\bar{y}_{h1} - \bar{y}_{h0})\sqrt{b_h}}{s_{den,h}}\right) \quad \text{with}$$

$$E\left[\hat{\theta}_{ptotal,h}\right] = \theta_{ptotal} \text{ and } \hat{\sigma}_{\hat{\theta}_{ptotal,h}} \approx \sqrt{\frac{s^2_{\bar{y}_{1h}-\bar{y}_{0h}}}{s^2_{den,h}} + \hat{\theta}^2_{ptotal,h}\left(1 - \frac{df_{s^2_{den,h}} - 2}{df_{s^2_{den,h}}\, c\left(df_{s^2_{den,h}}\right)^2}\right)} \quad (21)$$

Note that the adjustment White and Thomas [8] recommend for Hedges' g (Equation (5)) has been made to the estimated standard error in Equation (21).

Where $df_h > 2$, the expectation and sampling variance for the general unbiased SMD estimate are given by Huynh [12 p.6] and White and Thomas [8 p.143] respectively as θ_{SMD} and $\left[c(df_h)^2(df_h/df_h - 2)\left(s^2_{\bar{y}_{1h}-\bar{y}_{0h}}/s^2_{den,h} + \theta^2_{SMD}\right)\right] - \theta^2_{SMD}$. Given that the degrees of freedom are approximate so is the estimated standard error. It is worth noting that the accuracy of Satterthwaite [46] approximations may depend on the imprecision of the estimated component parameters. Equation (21) again simplifies. It is Hedges' g in

the standard case. It is Huynh's [12 p.6] h in his case. It is White and Thomas' [8 p.150] g_{adj} in their case and it is Hedges' [9 p.362] $DJ\left(b^2/c\right)$ in his.

The sampling distributions of the SMD estimators considered further in this paper are given in Table III. It can be seen that clustering and heteroscedasticity affect the trial SMD estimate and its standard error via the degrees of freedom, the trial SMD estimate via the denominator and its associated bias (where applicable), and finally, the standard error via $s^2_{\bar{y}_{1h}-\bar{y}_{0h}} \big/ s^2_{den,h}$.

[Insert Table III about here]

5. META-ANALYSIS METHODS USING THE IPD

Any IPD meta-analysis will require the systematic reviewer to obtain the trial datasets in which patients are linked to trials, interventions and outcomes. If clustering by care provider is also to be considered, provider identifiers linking patients to providers will additionally be required. Assuming all necessary data are available, the first step is to prepare the data for analysis and the second step is to fit the associated meta-analysis model.

5.1 Data Preparation

In preparing the data, Goldstein et al [24] suggest that it is necessary to standardise the outcome, giving it a common origin and metric. Firstly, by subtracting the mean in the control arm from observed patient-level outcomes, the outcomes within trials are given a common origin, transforming them to differences from this origin. This is important when trials use different measurement scales since standardised means, like absolute means, are expected to vary from trial to trial [24]; it is differences between standardised means that are assumed to be comparable. If measurement scales are the same across trials, the outcomes would already have a common origin, making this unnecessary [24]. Secondly, by dividing the differences by a common SD, outcomes are given a common metric. If the interpretation of an SMD is to be meaningful its metric should not be confounded with the mean differences within the trials (see Greenland [48] for a similar argument regarding standardised regression coefficients).

For this reason, the standardising metric must be common to all arms of a trial. In the standard case, where independence and common variance assumptions hold, the data would therefore be transformed, prior to analysis, as follows

$$y_{ikh}^{\text{Cohen's d}} = \frac{y_{ikh} - \bar{y}_{0h}}{s_{\bullet h}} \qquad (22)$$

where y_{ikh} is the outcome for patient i of treatment arm k of study h, $\bar{y}_{0h}$ is the mean in the control arm and $s_{\bullet h}$ the standardising metric from Equation (12).

Goldstein et al [24] assumed that the population value of the SD is known, and equal to the sample estimate, thereby ignoring Hedges' [15] small-sample bias. If all trials have large effective sample sizes, as in their example, this will have little impact, but as previously discussed it will lead to bias otherwise, even where the total sample size is large. This can be avoided by first dividing the metric by its correction factor $c\left(df_h\right)$ using Equation (4) as follows

$$y_{ikh}^{\text{Hedges' g}} = \frac{y_{ikh} - \bar{y}_{0h}}{s_{\bullet h} \Big/ c\left(df_{s_{\bullet h}^2}\right)} \qquad (23)$$

A similar transformation would be appropriate in the simple Behrens-Fisher situation. The difference here is that the divisor $s_{\bullet h}$ is now a linear combination of mean square terms so the degrees of freedom are not simply $n_{1h} + n_{0h} - 2$ as they are for Equation (23) but are taken from Table II.

In their example of studies of small versus large class size, where students are nested within classes, schools and studies, with schools crossed with interventions, Goldstein et al [24] suggested the following transformation,

$$y_{ijkh}^{\text{Crossed}} = \frac{y_{ijkh} - \bar{y}_{j0h}}{s_W} \qquad (24)$$

where y_{ijkh} is the outcome for patient i in school j of treatment arm k of study h, $\bar{y}_{j0h}$ is the cluster-level mean in the control arm and s_W the standardising metric. By standardising at the cluster-level, Goldstein et al [24] adopted a cluster-specific or

conditional approach. This is only possible because their studies had a crossed design (whereby schools had small and large class sizes) so absolute mean differences could be calculated at either the school or study level. Summary measures meta-analyses adopt a population-average or marginal approach, defining the origin at the trial-level. Defining the origin at the trial-level is necessary for nested designs as well since clusters relate only to one treatment arm. In line with their cluster-specific approach, Goldstein et al [24] used the within-cluster or level-1 SD as the metric within trials. They therefore implicitly assumed a random-intercept or random-coefficient model for the trials.

Where clusters are nested within interventions, as in our example, we suggest subtracting the marginal mean $\bar{y}_{0h}$ rather than the conditional mean $\bar{y}_{j0h}$ from y_{ijkh}. To be consistent with a two-level heteroscedastic model for the trials, we suggest as previously the relevant common metric is the pooled total SD within trials. Again, the small-sample bias can be avoided by dividing this metric by its correction factor $c(df_h)$ using Equation (4), with respect to the degrees of freedom in Table II. If the total SDs are used the transformation we suggest is simply,

$$y_{ijkh}^{\text{Pooled Total}} = c\left(df_{s^2_{T\bullet h}}\right)\left(\frac{y_{ijkh} - \bar{y}_{0h}}{s_{T\bullet h}}\right) \quad (25)$$

If the total mean squares are used, then it becomes,

$$y_{ijkh}^{\text{Pooled Naive}} = c\left(df_{s^2_{\bullet h}}\right)\sqrt{b_h}\left(\frac{\left(y_{ijkh} - \bar{y}_{0h}\right)}{s_{\bullet h}}\right) \quad (26)$$

5.2 Meta-Analysis Models

Once the outcome data have been transformed, the next step is to fit the appropriate model. Suppose y_i is the transformed outcome for the i^{th} patient, where $i = 1, \ldots, N$, and that it is normally-distributed. Suppose also that α_h represents the standardised mean outcome in the control arm of trial h (a fixed effect), and δ is a fixed treatment effect with K_i being an indicator variable for the intervention versus control arm. Using Goldstein's [49] notation for random effects, the standard fixed-effects meta-analysis model would be

$$y_i = \alpha_h + \delta K_i + e_i^{(1)} \qquad (27)$$

where and $e_i^{(1)}$ represents the level-1 random effect for patients and $e_i^{(1)} \sim N(0,1)$. We suggest that in the standard case (see Equations (22) and (23)), the following random-effects meta-analysis model should be fitted,

$$y_i = \alpha_h + \delta K_i + \tau_{trial(i)}^{(2)} K_i + e_i^{(1)} \qquad (28)$$

where $\tau_{trial(i)}^{(2)}$ represents the level-2 random effect mapping patients to trials and $\tau_{trial(i)}^{(2)} \sim N(0, \tau^2)$. Here, the trial effect is fixed but the treatment effect randomly varies across trials. In contrast to a fixed-effects meta-analysis model, Equation (28) respects the method by which data were standardised (accounting for the dependence induced by the data-driven transformation), defining the origin and metric at a trial-level. A fixed-effects meta-analysis model would be appropriate if outcomes were standardised across trials using $\bar{y}_0$ as the origin and $s_\bullet$ as the metric. However, defining the origin and metric at a meta-analysis level would only be appropriate if outcomes are standardised but measured with the same scales and standards across trials. This is relatively uncommon and not the case in our motivating example.

When the pooled within-treatment standard deviation $s_{\bullet h}$ is used in the context of the Behrens-Fisher problem, one possible parameterisation of the appropriate random-effects meta-analysis model is

$$y_i = \alpha_h + \delta K_i + \tau_{trial(i)}^{(2)} K_i + e_{i0}^{(1)}(1 - K_i) + e_{i1}^{(1)} K_i \qquad (29)$$

where e_{ik} are patient-level random errors for the control and treatment respectively and $e_{0l}^{(1)} \sim N(0, \sigma_{e0}^2)$ and $e_{1l}^{(1)} \sim N(0, \sigma_{e1}^2)$ with $(n_0 - 1)\sigma_{e0}^2 + (n_1 - 1)\sigma_{e1}^2 / n_0 + n_1 - 2 = 1$. Model (29) is appropriate because it respects the fact that the data were standardised with a pooled patient-level SD.

If, as described in Equations (25) and (26), the pooled total $s_{T\bullet h}$ or the pooled naïve $s_{\bullet h}$ SDs are used for meta-analysing fully-nested designs, we suggest extending Model (29) to allow for clustering within the trials under a common two-level heteroscedastic model for the trials, as follows

$$y_i = \alpha_h + \delta K_i + \tau^{(3)}_{trial(i)} K_i + u^{(2)}_{therapist(i)0}(1-K_i) + u^{(2)}_{therapist(i)1} K_i + e^{(1)}_{i0}(1-K_i) + e^{(1)}_{i1} K_i \quad (30)$$

where $u^{(2)}_{therapist(i)0}$ and $u^{(2)}_{therapist(i)1}$ are random intercepts for therapist j in the control and intervention arms, respectively, with $u^{(2)}_{0therapist(l)} \sim N(0, \sigma^2_{u0})$, $u^{(2)}_{1therapist(l)} \sim N(0, \sigma^2_{u1})$ and $\sigma_{u0u1} = 0$. Here, the level-2 and level-1 variances are allowed to differ between arms. It is the average of the total variances, i.e. $(n_0 - 1)(\sigma^2_{u0} + \sigma^2_{e0}) + (n_1 - 1)(\sigma^2_{u1} + \sigma^2_{e1})$ $/ n_0 + n_1 - 2$, that is equal to 1 now. It is important to note that Model (30) assumes the therapist ICC is equal across trials within arms. As we showed in our paper on the meta-analysis of absolute mean differences [25], this may be a strong assumption.

The assumption of between-trial homogeneity in the therapist ICC is more clearly respected when the data are standardised using the pooled naïve SD because the ICC used in the degrees of freedom and bias correction in Equation (26) can be taken from the pooled estimate across trials. When the pooled total SD is used directly (Equation (25)), an unstructured random structure is implicitly assumed. More complex models could be fitted, in theory, such as meta-regressions of the random parameters (see [25]), but for simplicity they were not considered here. If all trials are partially nested, $u^{(2)}_{therapist(i)0}$ is constrained to zero, and the term omitted from Model (30).

Where it is appropriate to assume a common random intercept model across trials, as may be the case for cluster randomised trials, Model (30) simplifies to

$$y_i = \alpha_h + \delta K_i + \tau^{(3)}_{trial(i)} K_i + u^{(2)}_{cluster(i)} + e^{(1)}_i \quad (31)$$

6. APPLICATION TO THE MOTIVATING EXAMPLE

Short-term outcomes relating to the BDI [5], the GHQ [36], the HADS-D [3] and the Symptom Index were available for 850 patients from seven [29-35] counselling in primary care trials. Of these, 494 (58%) were allocated counselling with one of 56 counsellors. Overall, the cluster sizes ranged from 1 to 47, with a median of 4.5 and an interquartile range of 2-10.5. Data were available for five or more patients for 33 of the counsellors. Table IV gives descriptive statistics for the seven included trials. The total mean squares (i.e. the pooled naïve variance) and the pooled total variance estimates are similar indicating the bias arising from using the pooled naïve SD to estimate the pooled total SD is likely to be minimal here. The published meta-analysis used a slightly different subset of patients as we excluded 18 patients with missing counsellor identifiers from all analyses.

[Insert Table IV about here]

6.1 Summary-Data Meta-Analyses

Sample estimates of parameters used in estimating the SMDs are given in Table V. ANOVA estimates of the counsellor ICC range from -0.14 to 0.29. The possibility of negative ICC estimates arises as ANOVA estimation is consistent with a common correlation model rather than a variance components model [50]. By definition, the lower bound on the ICC is zero for a variance components model because a between-cluster variance cannot be negative. It is the design effect that cannot be negative in ANOVA estimation. Design effects based on the raw ICCs range from 0.20 for Hemmings [33] to 1.73 for Harvey [32]. The negative ICCs found here are likely to be a consequence of sampling error arising from limited counsellors per trial and a small population ICC. As no evidence of heterogeneity was found in the ICCs between trials [51], to simplify our summary-data meta-analyses, we assumed a common counsellor ICC of 0.022, using a weighted average of these throughout, regardless of the model. We will return to this assumption in the discussion. Using this assumed fixed ICC, design effects vary from 1.04 for Chilvers [30] to 1.80 for Hemmings [33]. To reflect a general lack of knowledge about cluster size distributions, we assumed

equal cluster sizes as well. This assumption is questionable, as a few counsellors were responsible for the majority of counselling in Chilvers [30] and King [34].

[Insert Table V about here]

Table V gives sample estimates for Satterthwaite degrees of freedom, Hedges small-sample bias and the bias associated with the choice of estimator for the standardising metric for each case and for each included trial. The degrees of freedom drop for the Behrens-Fisher, pooled total and pooled naïve cases compared to the standard Hedges g. However this has limited impact, with the correction factor for the small-sample bias, c(df), being close to one and generally unaffected by the choice of model. As we expected, comparing the total mean squares and pooled total variance, the correction factor for using the naïve SD is also close to one: it is precisely one in all other cases.

Table VI gives summary-data estimates and associated standard errors for fixed- and random-effects meta-analyses of SMDs based on four standardising metrics: Hedges g is the standard case against which the Behrens-Fisher, pooled total and pooled naïve cases are compared. We also present the impact of iterating estimates of the standard errors for the weights, giving the estimates and associated standard errors from both the initial (i.e. not iterated) and iterated models. Using the reduced dataset and fixed-effects meta-analysis, the initial standard pooled SMD was estimated to be -0.24 (SE = 0.013; 95% CI -0.27 to -0.22) while the iterated equivalent was -0.26 (SE = 0.072; 95% CI -0.40 to -0.12), similar to the published result (SMD= -0.24, 95% CI -0.38 to -0.10). This highlights the importance of iterating here, as the standard error of the pooled SMD is underestimated initially, leading to over-precise confidence intervals.

[Insert Table VI about here]

The pooled SMD and its associated standard error are similar for the fixed-effects models for all four SMDs, increasing only slightly for the Behrens-Fisher, pooled total and pooled naïve cases compared to the standard case. The recommended model in each case is a random-effects meta-analysis model. Initial and iterated estimates are much more similar here. The pooled SMD and its associated standard error are also

almost identical across the four cases as well. This is because the between-trial variance dominates. A possible explanation for this is differences in the counselling provided and in the patients included across trials.

One reason for the limited impact of accounting for treatment-related clustering in the summary-data meta-analyses could be the size of the ICC assumed. So we performed a sensitivity analysis, increasing the assumed ICC in analyses reported in Table VI. The conclusions remain unchanged. However, as expected and contrary to trial estimates of absolute mean differences [25], trial estimates of SMDs were pulled towards the pooled treatment effect estimate as the ICC increased, although not perceptibly in the range of the ICC expected here. Random-effects meta-analysis estimates remained more stable than their fixed-effects counterparts, providing further support for the conclusion that treatment-related clustering has more impact on fixed-effects than on random-effects meta-analyses. The DerSimonian-Laird estimate of the between-trial variance increased until the ICC was in mid-range, decreasing again up to its maximum, illustrating that the total and naïve SMDs are a function of the ICC. The cluster sizes were known in our example. If the cluster sizes had been assumed, further analyses would be recommended to assess the sensitivity of the conclusions to these assumptions.

6.2 IPD Meta-Analyses

In contrast to summary-data meta-analyses, those based on IPD make it practical to relax assumptions relating to the cluster size distribution. They also make it clearer what is being assumed. Data were prepared separately for each case using the relevant degrees of freedom and correction factors for (i) the small-sample bias and (ii) bias associated with the choice of estimator for the standardising metric using Equations (25) and (26). IPD models were implemented using Restricted Maximum Likelihood (REML) [49] with the `mixed` command in Stata 13 (see Table VII) and with Restricted Iterative Generalised Least Squares (RIGLS) [49] in MLwiN. Both gave comparable results. Details of programming code for Stata are given as supporting web materials.

[Insert Table VII about here]

Table VII gives IPD estimates and associated standard errors for fixed- and random-effects meta-analyses of SMDs based on the four standardising metrics as before. The IPD counterparts to the summary-data estimates reported in Table VI are very similar (Table VIII gives a summary of the summary-data and IPD results), with the standard pooled SMD from the fixed-effects meta-analysis being -0.261 (SE = 0.071; 95% CI -0.40 to -0.12). This indicates that little is gained by using a full-likelihood approach and accounting for variability in cluster sizes here. The patient-level variance estimate is 0.992. To interpret the SMD in SD units, this should be precisely equal to one. That it is not suggests the SMD is slightly under estimated, as -0.261/0.992 = -0.263. The same can be said for the standard random-effects estimate, where Model (28) applies: here -0.263/0.987 = -0.266.

[Insert Table VIII about here]

In the Behrens-Fisher case, the appropriate model explicitly allows for between-arm heteroscedasticity at the patient-level. The relevant SMD is the extension of Huynh's [12] SMD that allows for a ratio of sample sizes between arms other than one. Boot et al [29], Harvey et al [32] and Hemmings [33] all had unequal sample sizes favouring the counselling arm, making this issue pertinent to this example. Under the random-effects model given in Model (29), the pooled SMD was estimated to be -0.263 (SE = 0.093) with the IPD. In this case, it is the average of the patient-level variances i.e. ((494-1)*0.867+(356-1)*1.153)/(850-2)=0.987 that defines the metric. This again implies that the SMD is slightly under-estimated, as -0.263/0.987=-0.266.

In the pooled total and pooled naïve cases, the appropriate model explicitly allows for between-arm heteroscedasticity at the counsellor- and at the patient-levels. Estimates from the random-effects meta-analysis model given in Model (30) are identical, with the pooled SMD estimated to be -0.264 (SE = 0.092). Under this model, the metric is ((494-1)*(0.842+0.035)+(356-1)*1.154)/(850-2)=0.993 for the pooled total case and ((494-1)*(0.841+0.037)+(356-1)*1.151)/(850-2)=0.992 for the pooled naïve case. It is therefore close to one, with the final SMD estimated to be -0.264/0.993=-0.266 (SE =

0.092; 95% CI -0.45 to -0.09) in favour of the counselling arm. This indicates that counselling reduces short-term mental health symptoms by an average of 0.266 SDs compared to no counselling and that this reduction remains statistically significant at the 5% level. According to Cohen [13], an effect size of 0.2 equates to a small effect. The confidence interval is wide including moderate effect sizes as well as trivial ones. Heterogeneity in the size of effects between counsellors and trials suggests that more could be done to optimise counselling in primary care.

7. DISCUSSION

As we have highlighted, the meta-analysis of SMDs is more complicated than that of absolute mean differences (see [25]), especially where clustering associated with care providers is probable. This is partly because an SMD is a ratio and partly because its denominator is also estimated. This leads to the data-driven transformations seen in the IPD case. So, in contrast to the meta-analysis of absolute mean differences, use of a fixed effects meta-analysis model is less defensible, summary statistics are biased in small samples [6, 15], their sampling variance depends on the population parameter (see Equation (4)) and their sampling distribution follows a non-central t-distribution (Equation (18)). These are true of all meta-analyses of SMDs. Where there is between-arm heteroscedasticity in provider and patient level variances, the size of the SMD, its small-sample bias, its sampling variance and interpretation additionally depend on the choice of standardising metric.

A general approach has been described which allows for treatment-related clustering in the meta-analysis of normally-distributed outcomes from randomised trials with two-level nested designs. Building on the work of Hedges [9], Huynh [12], Goldstein et al [24] and White and Thomas [8], we have recommended a pooled total SD as the standardising metric, using the pooled naïve SD to estimate this where a pooled total SD is not available in trial reports. The advantages of the pooled total SMD are that i) it is general, in the sense that it encompasses Hedges', Huynh's and White and Thomas' estimators as special cases, allowing the assumptions of independence and common variance to simultaneously be relaxed within trials but also their sample

sizes to differ across arms, ii) it can be estimated using the pooled naïve SD where published data is limited and iii) its interpretation is comparable across trial designs, allowing extensions for pooling mixed trial designs.

In our example, all trials had partially-nested designs, so that the counsellor variance was equal to zero in the control arm. Some of the trials also had unequal sample sizes across arms. As a result, the methods described by White and Thomas [8] and Hedges [9] could not be applied. In the context of the IPD, a random-intercept model could have been assumed for the trials, but a choice would have had to be made between including patients in the control arm as clusters of size one or as clusters of size n_{h0}. If clusters of size one were used, the within-cluster SD would not be defined for the control arm and would be estimated solely within the treatment arm; the between-cluster SD would be available in both arms, but it would unlikely be equal. If clusters of size n_{h0} were used, the between-cluster SD would not be defined in the control arm and would be estimated solely from the treatment arm. Also while the within-cluster SD is available in both arms, the number of clusters is unequal, giving greater weight to the treatment arm. In neither case is a random-intercept model appropriate.

We could have extended Glass' SMD [11], using the control arm SD as the metric of choice. This option was initially quite appealing for our example as the point estimate would be independent of treatment-related clustering, minimising impact of between-trial heterogeneity in the ICC. The drawback became clear when the corresponding IPD meta-analysis model was considered. If the control arm SD is used as the metric then the denominator of the ICC is the patient SD in the control, not treatment, arm. This mis-specifies the variance-covariance structure of the two-level heteroscedastic model assumed for the trials and makes interpretation of random effects less straight-forward. The proposed SMD based on the pooled total SD addresses these limitations.

Using our proposed metric for our example of counselling in primary care, we found that the impact of treatment-related clustering on the pooled SMD estimate and its standard error was not important. Our sensitivity analyses for the meta-analysis

published by Bower and Rowland [28] did not change their conclusions. In hindsight, the reasons for this are obvious: (i) the ICC and the cluster sizes were both small, so the variance inflation factors were small, (ii) a larger number of patients were allocated to counselling compared to no counselling, and (iii) the between-trial heterogeneity in the treatment effect was dominant. It is therefore unsurprising that the conclusions of the published meta-analysis remain unchanged.

However, on top of this, in our summary-data meta-analyses, we assumed the population counsellor ICCs were the same across trials, and equal to our pooled estimate. Making this assumption simplified our analyses but it is clearly a limitation of our approach. It was motivated by separately finding no evidence of heterogeneity [51]. Further work is needed to explore the implications of allowing the ICCs to vary across trials, and of using (truncated) ICC estimates in such analyses, where this assumption is not reasonable. Although it is unlikely that the conclusions of the meta-analysis would change if a more complex summary-data model had been fitted, it is possible that this contributed to the limited impact of treatment-related clustering observed here. Impact may not be limited in general, however. While treatment-related clustering has historically rarely been taken into account in trials, it is now recommended that it is considered in trials of non-pharmacological treatments [52]. Similar broad guidance is made by Cochrane [53] with regard to meta-analyses, although the methods are just becoming available. This paper contributes to the literature supporting the uptake of this guidance.

A key assumption made in all meta-analyses of SMDs is that patient outcomes are normally-distributed within trials. This assumption allows the standardising metric to be distributed proportionally to a chi-square with known degrees of freedom in the standard case, and with degrees of freedom given using a Satterthwaite approximation otherwise. Further work is needed to explore the impact of departures from normality. Alternative approaches might also be investigated, such as the use of robust estimates of the variance of the standardising metric, particularly for small samples. Another assumption is that Hedges' small-sample bias should be corrected for using Hedges' g instead of Cohen's d. In moderate to large samples this will be unimportant, but it is

the effective sample size rather than the number of observations that determines the impact of this bias so where the ICC or cluster size is large, Hedges' correction may still be important.

Our IPD meta-analyses start to show how the model chosen depends on the choice of metric. Specifically we explored the relationship between the choice of metric and the model that preserves the interpretation of that metric. We found the metric implied by each model was not precisely what we expected, which implies that there is a further bias not identified here. In the standard case, we suggest that a random-effects meta-analysis model is appropriate since the SD estimate tends to be trial-specific. For the Behrens-Fisher case, we suggest the metric and model should reflect heterogeneity in the patient SD across arms. For nested designs with treatment-related clustering, we propose that the patient and cluster-level SDs should be allowed to vary across arms. Where this was so, we expected the relevant SD estimated from the model to equal one, so the SMD can simply be interpreted as a mean difference given in SD units and the counsellor variance directly estimates the counsellor ICC. Two explanations for this disparity could be explored further. The first relates to the relative weighting of data by the standardising SD and REML, the second to between-trial heterogeneity in counsellor ICCs affecting the standardising metric. Secondly, it also became clear that the model depends on the level at which data is standardised when contrasting our models with those proposed by Goldstein [24]. A population-average model is arguably more appropriate for meta-analysing nested designs. More work is needed to investigate the implications of this in meta-analyses incorporating treatment-related clustering, generalising the methods proposed by Bohning et al [54] and Viechtbauer [55]. Thirdly, it may not always be safe to assume a common origin across trials. If a random trial intercept was included in an IPD model, correlation can be estimated between heterogeneity in a SMD and its origin. While this is regarded as a nuisance in summary-data meta-analyses, it may be worth considering in IPD meta-analyses. Further work is needed to understand these issues more, and the potential biases associated with them.

The focus of this paper has been on meta-analyses of SMDs where all included trials are not only addressing the same research question, but also have comparable designs. In our motivating example, all seven trials had a partially nested design. It would be straightforward to extend the methods outlined here to situations where all the trials have fully nested designs. In both cases, the assumption of between-trial homogeneity in the random effects is arguably tenable. We have previously argued [25] that this is not the case for meta-analyses of mixed clustered trial designs. There, we argue ICCs vary not only between trial arms but also between trial designs. An implication of this for meta-analyses of SMDs is that the metric would vary systematically between trials with different designs. It becomes crucial that a metric has comparable interpretations across trial designs so variation in the size of the standardising metric is a reflection only of the use of different trial designs to address a common research question. It is important that these issues are considered if mixed trial designs are to be included in a meta-analysis: the general metric we have proposed is only half the story. Models that preserve the interpretation of that metric in different situations are also needed.

In our motivating example, there was the potential for clustering by the GP. GP care tended to be a co-intervention, with GPs crossed with treatment arms. As GPs were not blinded, an interaction between GPs and treatment arm is plausible. Information on GP involvement was limited, however, with GP identifiers only recorded in one or two of the trials. We recommend future trials record identifiers for all significant care providers, whether they are delivering the trial intervention or not. This will enable meta-analysts to incorporate, or explore incorporating, trials with multiple therapist-per-patient designs extending [56]. Our motivating example also included additional levels, in that repeated measures were available over time. As the follow-up periods included as "short-term outcomes" ranged from 6 weeks to 6 months, and most of the trials included more than one outcome visit, further work is planned to fit realistic meta-analysis models to the IPD available, building on this methodological work and that of others (e.g. [57] and [58]), aimed more at a clinical audience.

None of the seven trials we included had accounted for treatment-related clustering in their published analyses. This made it important to take account of treatment-related clustering in our meta-analysis. If they had done so appropriately, then we would not

have needed to take additional account of it in our summary-data meta-analyses as the trial estimates and their standard errors could be pooled directly. IPD meta-analyses would remain as outlined however. As we had access to all of the relevant IPD, we had maximum flexibility. As we have shown here, the pooled treatment effect and its 95% confidence interval are very similar for summary-data and IPD meta-analyses. Beyond standard access to sample sizes, means and SDs by trial arm, we assumed access to average cluster sizes and a range of realistic ICCs by trial arm. Based on our experience with our motivating example, these are likely to be readily available from trial reports and the more general literature (e.g. [59]). Use of the total mean squares (or naïve SD) typically reported is possible in place of the pooled total SD as we have described.

In conclusion, in the presence of treatment-related clustering, meta-analysis of SMDs is more complicated than that of absolute mean differences, and hence more difficult to interpret, but it is possible if sufficient care is taken using the methods described here and extensions to these. Specific guidance is needed in the Cochrane Handbook to facilitate the uptake of these methods. The code used to program them in Stata is available from the first author on request.

REFERENCES

1. Deeks, J.J., D.G. Altman, and M.J. Bradburn, Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis, in Systematic Reviews in Health Care: Meta-analysis in Context, M. Egger, G. Davey-Smith, and D.G. Altman, Editors. 2001, BMJ Books: London. p. 285–312.
2. Glass, G.V., Primary, secondary and meta-analysis of research. Educational Researcher 1976. **5**: p. 3-8.
3. Zigmond, A.S. and R.P. Snaith, The hospital anxiety and depression scale Acta Psychiatrica Scandinavica, 1983. **67**(6): p. 361–370.
4. Kroenke, K., R.L. Spitzer, and J.B.W. Williams, The PHQ-9: Validity of a Brief Depression Severity Measure. Journal of General Internal Medicine, 2001. **16**(9): p. 606–613.
5. Beck, A.T., et al., An inventory for measuring depression. Archives of General Psychiatry, 1961. **4**(6): p. 561-571.
6. Hedges, L.V., Estimation of effect size from a series of independent experiments. Psychological Bulletin, 1982. **92**(2): p. 490-499.

7. Donner, A. and N. Klar, Design and Analysis of Cluster Randomization Trials in Health Research 2000, London: Arnold.
8. White, I.R. and J. Thomas, Standardized mean differences in individually-randomized and cluster-randomized trials, with applications to meta-analysis. Clinical Trials, 2005. **2**: p. 141–151.
9. Hedges, L.V., Effect sizes in cluster randomized designs. Journal of Educational and Behavioral Statistics, 2007. **32**: p. 341–370.
10. Kim, S.-H. and A.S. Cohen, On the Behrens-Fisher problem: A review. Journal of Educational and Behavioral Statistics 1998. **23**(4): p. 356-377.
11. McGaw, B. and G.V. Glass, Choice of the metric for effect size in meta-analysis. American Educational Research Journal 1980. **17**(3): p. 325-337.
12. Huynh, C.L. A unified approach to the estimation of effect size in meta-analysis. in Annual Meeting of the American Educational Research Association. 1989.
13. Cohen, J., Statistical Power Analysis for the Behavioral Sciences. 1988, New York: Academic Press.
14. Grissom, R.J. and J.J. Kim, Review of assumptions and problems in appropriate conceptualization of effect size. Psychological Methods, 2001. **6**: p. 135-146.
15. Hedges, L.V., Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics, 1981. **6**(2): p. 107-128.
16. Morris, S.B. Effect size estimation from pretest-posttest-control designs with heterogeneous variances. in 20th Annual Conference of the Society for Industrial and Organizational Psychology. 2005.
17. Morris, S.B., Estimating effect sizes from pretrest-posttest-control group designs. Organizational Research Methods, 2008. **11**(2): p. 364-386.
18. Morris, S.B. and R.P. DeShon, Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. Psychological Methods, 2002. **7**(1): p. 105-125.
19. Becker, B.J., Synthesizing standardized mean-change measures. British Journal of Mathematical and Statistical Psychology, 1988. **41**: p. 257-278.
20. Carlson, K.D. and F.L. Schmidt, Impact of experimental design on effect size: Findings from the research literature on training. Journal of Applied Psychology, 1999. **84**: p. 851-862.
21. Marshall, M., et al., Systematic reviews of the effectiveness of day care for people with severe mental disorders: (1) acute day hospital versus admission; (2) vocational rehabilitation; (3) day hospital versus outpatient care. Health Technology Assessment 2001. **5**(21): p. 1-75.
22. Birks, J. and L. Flicker, Selegiline for Alzheimer's disease. Cochrane Database of Systematic Reviews, Issue 1. Art. No.: CD000442, 2003.
23. Trialists, E.S.D., Services for reducing duration of hospital care for acute stroke patients. Cochrane Database of Systematic Reviews, Issue 2. Art. No.: CD000443., 2005.
24. Goldstein, H., et al., Meta-analysis using multilevel models with an application to the study of class size effects. Applied Statistics, 2000. **49**(3): p. 399-412.

25. Walwyn, R. and C. Roberts, Meta-analysis of absolute mean differences from randomised trials with treatment-related clustering associated with care providers. Statistics in Medicine, 2015. **34**: p. 966–983.
26. Walwyn, R. and C. Roberts, Therapist variation within randomised trials of psychotherapy: implications for precision, internal and external validity. Statistical Methods in Medical Research, 2010. **19**: p. 291–315.
27. Roberts, C. and S.A. Roberts, Design and analysis of clinical trials with clustering effects due to treatment. Clinical Trials 2005. **2**: p. 152–162.
28. Bower, P. and N. Rowland, Effectiveness and cost effectiveness of counselling in primary care. . . Cochrane Database of Systematic Reviews 2006; 3. Art. No.: CD001025, 2006.
29. Boot, D., et al., Evaluation of the short-term impact of counseling in general practice. Patient Education & Counseling, 1994. **24**(1): p. 79-89.
30. Chilvers, C., et al., Antidepressant drugs and generic counselling for treatment of major depression in primary care: Randomised trial with patient preference arms. British Medical Journal 2001. **322**(7289): p. 772-775.
31. Friedli, K., et al., Randomised controlled assessment of non-directive psychotherapy versus routine general-practitioner care. Lancet 1997. **350**(9092): p. 1662-1665.
32. Harvey, I., et al., A randomized controlled trial and economic evaluation of counselling in primary care. British Journal of General Practice, 1998. **48**(428): p. 1043-1048.
33. Hemmings, A., Counselling in primary care: A randomised controlled trial. Patient Education & Counseling, 1997. **32**(3): p. 219-230.
34. King, M.B., et al., Randomised controlled trial of non-directive counselling, cognitive-behaviour therapy and usual general practitioner care in the management of depression as well as mixed anxiety and depression in primary care. Health Technology Assessment, 2000. **4**(19): p. 1-83.
35. Simpson, S., et al., A randomised controlled trial to evaluate the effectiveness and cost-effectiveness of counselling patients with chronic depression. Health Technology Assessment 2000. **4**(36).
36. Goldberg, D., Manual of the General Health Questionnaire. 1978, Windsor: NFER-Nelson.
37. Bower, P., N. Rowland, and R. Hardy, The clinical effectiveness of counselling in primary care: a systematic review and meta-analysis. Psychological Medicine, 2003. **33**: p. 203–215.
38. Hedges, L.V., A random effects model for effect sizes. Psychological Bulletin, 1983. **93**(2): p. 388-395.
39. Whitehead, A., Meta-Analysis of Controlled Clinical Trials. 2002, New York: Wiley.
40. Thompson, S.G. and J.P.T. Higgins, How should meta-regression analyses be undertaken and interpreted? Statistics in Medicine, 2002. **21**(11): p. 1559–1573.
41. van Houwelingen, H.C., L.R. Arends, and T. Stijnen, Advanced methods in meta-analysis: multivariate approach and meta-regression. Statistics in Medicine, 2002. **21**(4): p. 589-624.
42. Birge, R.T., The calculation of errors by the method of least squares

Physical Review, 1932. **16**: p. 1–32.
43. Cochran, W.G., Problems arising in the analysis of a series of similar experiments. Journal of the Royal Statistical Society, 1937. **4**(Supplement): p. 102–118.
44. DerSimonian, R. and N.M. Laird, Meta-analysis in clinical trials. Controlled Clinical Trials 1986. **7**: p. 177–188.
45. Sidik, K. and J.N. Jonkman, Robust variance estimation for random effects meta-analysis. Computational Statistics & Data Analysis, 2006. **50**: p. 3681–3701.
46. Satterthwaite, F.E., An approximate distribution of estimates of variance components. Biometrics Bulletin, 1946. **2**(6): p. 110-114.
47. Box, G.E.P., Some theorems on quadratic forms applied to the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. Annals of Mathematical Statistics, 1954. **25**: p. 290-302.
48. Greenland, S., J.J. Schlesselman, and M.H. Criqui, The fallacy of employing standardized regression coefficients and correlations as measures of effect. American Journal of Epidemiology, 1986. **123**(2): p. 203-208.
49. Goldstein, H., Multilevel Statistical Models. 3rd Edition ed. 2003, London: Arnold.
50. Wang, C.S., B.S. Yandell, and J.J. Rutledge, The dilemma of negative analysis of variance estimators of intraclass correlation. Theoretical and Applied Genetics, 1992. **85**: p. 79–88.
51. Walwyn, R., Therapist variation within meta-analyses of psychotherapy trials 2010, University of Manchester: Manchester, UK.
52. Boutron, I., et al., Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: explanation and elaboration. Annals of internal medicine, 2008. **148**(4): p. 295-309.
53. Higgins, J.P.T. and S. Green, Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011], 2011, The Cochrane Collaboration.
54. Bohning, D., et al., Some general points in estimating heterogeneity variance with the DerSimonian-Laird estimator. Biostatistics, 2002. **3**(4): p. 445–457.
55. Viechtbauer, W., Bias and efficiency of meta-analytic variance estimators in the random-effects model. Journal of Educational and Behavioral Statistics, 2005. **30**: p. 261–293.
56. Roberts, C. and R. Walwyn, Design and analysis of non-pharmacological treatment trials with multiple therapists per patient. Statistics in Medicine, 2013. **32**: p. 81–98.
57. Jones, A.P., et al., Meta-analysis of individual patient data versus aggregate data from longitudinal clinical trials. Clinical Trials, 2009. **6**: p. 16–27.
58. Trikalinos, T.A. and I. Olkin, Meta-analysis of effect sizes reported at multiple time points: A multivariate approach. Clinical Trials, 2012. **9**: p. 610–620.
59. Baldwin, S.A., et al., Intraclass correlation associated with therapists: Estimates and applications in planning psychotherapy research. Cognitive Behaviour Therapy, 2011. **4**(1): p. 15–33.