Article:
Walwyn, R and Roberts, C (2015) Meta-analysis of absolute mean differences from randomised trials with treatment-related clustering associated with care providers. Statistics in Medicine, 34 (6). pp. 966-983. ISSN 0277-6715

https://doi.org/10.1002/sim.6379

Full Title:

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

Short Title:

Meta-analysis of mean differences from trials with clustering effects

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: mean difference; 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. Rebecca Walwyn and Chris Roberts are members of the UK Mental Health Research Network (MHRN) Methodology Research Group.

Abstract

Nesting of patients within care providers in trials of physical and talking therapies creates an additional level within the design. The statistical implications of this are analogous to those of cluster-randomised trials, except that the clustering effect may interact with treatment and can be restricted to one or more of the arms. The statistical model that is recommended at the trial-level includes a random effect for the care provider, but allows the provider and patient level variances to differ across arms. Evidence suggests that, while potentially important, such within-trial clustering effects have rarely been taken into account in trials and do not appear to have been considered in meta-analyses of these trials.

This paper describes summary measures and individual-patient-data (IPD) methods for meta-analysing absolute mean differences from randomised trials with two-level nested clustering effects, contrasting fixed and random effects meta-analysis models. It extends methods for incorporating trials with unequal variances and homogeneous clustering to allow for between-arm and between-trial heterogeneity in ICC estimates. The work is motivated by a meta-analysis of trials of counselling in primary care, where the control is no counselling and the outcome is the Beck Depression Inventory (BDI). Assuming equal counsellor ICCs across trials, the recommended random-effects heteroscedastic model gave a pooled absolute mean difference of -2.53 (95% CI -5.33 to 0.27) using summary measures and -2.51 (95% CI -5.35 to 0.33) with the IPD. Pooled estimates were consistently below a minimally important clinical difference of 4 to 5 points on the BDI.

1. INTRODUCTION

Where the treatment a patient receives is delivered by a health professional, such as in talking or physical therapies or surgery, patient outcomes may vary systematically by care provider. Variation between clusters, or, in this case, care providers, leads to correlation among patient outcomes within clusters, thereby violating the assumption of independence on which standard methods of analysis are based. Such correlation arises when care providers differ in characteristics related to outcome, such as training, skill, experience or empathy. The usual situation in psychotherapy is that treatment is provided by different samples of clusters in each arm in what will be referred to as a nested therapist design (patients are allocated to care providers within treatments). As this is a special case of the more generic fully-nested design (where clusters formed at recruitment, treatment or outcome assessment are nested within treatments), the statistical implications of provider clustering in nested therapist designs are analogous to the implications of recruitment-related clustering in standard cluster randomised trials, in which clusters are randomly allocated to treatments. The latter are now widely recognised [1]. Ignoring provider clustering can also result in treatment estimates that are too precise and standard errors that are too small. There are also crossed designs in which all treatments are provided in each cluster so that the clusters and treatments are crossed. This covers a cluster randomised crossover design [2-4] in which sequences of treatments are randomised to clusters as well as a crossed therapist design in which patients are allocated to treatments within care providers (see Walwyn and Roberts [5] for further details).

Cluster randomised trials often assume that the clustering effect is homogeneous across treatment arms, so a random intercept model is appropriate and a single intra-class correlation coefficient (ICC) is estimated. Care provider clustering may be treatment-specific, however, in that provider characteristics may differ across arms, for instance with greater skill or different training being required for one therapy compared to another. There may also be greater standardisation of one therapy, or one may be more established so that there is greater experience associated with it. Between-arm heterogeneity in the clustering effect, or treatment-related clustering, complicates matters so methods outlined for cluster randomised trials need to be extended for therapist designs. The statistical model that is recommended for nested therapist designs [6] includes a random effect for the care provider but allows the provider and patient level variances to differ across arms. We refer to this as a two-level heteroscedastic model [5]. As such, a separate ICC is estimated in each treatment arm. For

crossed therapist designs, the recommended model [5] is a random coefficient model, which includes a random intercept for the care provider but also allows the treatment effect to vary across care providers. In this case, between-provider variation in outcome increases precision of the treatment effect while between-provider variation in treatment effects decreases it. In the situation where clustering is absent from one arm, for example where the control is a waitlist or no treatment, the design is referred to as partially-nested or partially-crossed [5]. In this case, the between-cluster variance is constrained to zero in the no clustering arm. Incorporating crossed designs into meta-analyses raises different issues. These are beyond the scope of this paper and so will not be considered further here.

Care provider variation has widespread implications for the design and analysis of trials with nested designs. It affects not only the precision of treatment effect estimates [5-8] but also their internal and external validity [5, 6, 8]. It is now accepted that it needs to be considered in trials of non-pharmacological treatments [9]. However, a yet unpublished systematic methodological review of Cochrane reviews of comparative studies involving psychotherapy found that, while potentially important, such within-trial variation has rarely been taken into account in psychotherapy trials and does not appear to have been considered in meta-analyses of these trials [10]. Statistical pooling or meta-analysis of summary-data across trials can be viewed as a two-stage process in which summary statistics are first extracted from each trial and then a weighted average is calculated of them [11-12]. Where outcomes are normally distributed, the summary statistic for the treatment effect may be an absolute or standardised mean difference. Our methodological review included 101 Cochrane reviews and 1816 unique studies, 1345 of which involved psychotherapy given by care providers. Similar issues would apply to meta-analyses of surgical or educational interventions, physiotherapy, occupational therapy or speech therapy, where nested trial designs have been used. Where a published trial analysis has adequately allowed for the care provider, there would be no need to make further allowance in a summary-data meta-analysis. Problems only arise where no allowance has been made at the trial-level or where an inappropriate model has been used. In the current context it is likely that no allowance will have been made at the trial-level. As such, the problems outlined in this paper are expected to be quite common in practice.

The past decade has seen growing interest in the specific methodological challenges faced in the meta-analysis of randomised trials with correlated data. Methods have been proposed for pooling trials with repeated-measures [13-16], for crossover trials [17-20], and for cluster-

randomised designs [21-26]. What is common across this literature is a consideration of the impact of within-trial clustering when combining data from trials with complex data structures, particularly where this has been ignored in published trial analyses. Drawing on this literature, the Cochrane Handbook [27] cites methods for the meta-analysis of cluster-randomised and crossover trials. It briefly mentions clustering in individually randomised trials arising from health professionals but gives no specific guidance beyond stating that the issues are similar to those in cluster randomised trials, citing Lee and Thompson [7]. It makes no mention of treatment-related clustering effects, which may arise in individually-, or indeed in cluster-, randomised trials where interventions are delivered by care providers.

The presence of between-trial heterogeneity in ICCs for care providers raises further issues not previously considered in the literature. This heterogeneity might arise from disparities in the cluster or patient level variances across trials. Possible causes could be differences in the level of treatment standardisation or patient eligibility criteria between trials. One option would be to estimate separate ICCs in each trial for each arm. An alternative might be to estimate a single ICC across trials for each arm. Here, treatment-specific ICCs are pooled across trials. A further option would be to adopt a middle road and investigate the use of meta-regression models for the variance parameters. This paper considers methods for meta-analysing absolute mean differences from individually-randomised trials with two-level nested designs and treatment-related clustering. Both fixed- and random-effects meta-analysis models are considered, along with both summary-data and individual-patient-data (IPD) approaches. As with any meta-analysis of absolute mean differences using a summary-data approach, the sample means, variances and sizes are needed in each trial arm. To implement the methods described here, the ICC and average cluster size are also required for each trial by arm. The IPD approach assumes researchers have collected cluster identifiers, linking clusters to participants. The feasibility of obtaining these is commented on in the discussion.

We begin in section 2 by outlining the example that motivated this work. In section 3 we go on to review the recommended model at the trial level for fully and partially nested therapist designs. We then extend standard summary-data and IPD approaches to the meta-analysis of absolute mean differences in sections 4 and 5, respectively, outlining meta-regression models in section 6 and illustrating the proposed methods with our example in section 7. Section 8 contains a discussion, including limitations. Focusing initially on absolute mean differences has several advantages. Firstly, their large-sample estimates are unbiased, their sampling

variances are independent of the population parameter, and their sampling distribution is normally distributed [28]. As this is not the case for standardised mean differences, this avoids some of the added complications encountered when pooling the latter, allowing the general implications to be considered first. A separate paper, drawing on earlier work [10], is currently in preparation focusing on problems associated with pooling standardised mean differences in this context.

2. MOTIVATING EXAMPLE

The main point of contact for patients presenting in primary care in the UK is their general practitioner (GP) and associated primary care team. One in three is estimated to be affected by mental health problems [29]. The case for providing psychological therapies, including counselling, within the NHS has been made, with a rapid rise in counselling in primary care seen since 1990. Half the general practices in England were estimated to have a counsellor attached by 2000 [30]. The background of counsellors working in this setting is variable [31]. Counselling is typically brief, usually involving 6 to 10 sessions, each of 50 minutes [32]. The counselling process is characterised by three stages, operating by means of the relationship between the counsellor and the patient [31]. The focus is initially on building trust. The counsellor encourages the patient to describe the situation that is affecting them and makes a systematic assessment. The emphasis then turns to creating changes which give the patient additional resources they can subsequently draw upon. The way this is done depends on the theoretical model the counsellor is applying. Finally, alternative means of using the resources are considered, put into action and reflected upon. It is usual for counsellors to apply eclectic therapeutic approaches for a wide range of social and clinical problems.

Bower and Rowland [33] published a systematic review and meta-analysis of the clinical and cost-effectiveness of counselling in primary care, including eight trials. The largest meta-analysis compared counselling plus GP care to GP care alone, using the short-term outcomes measuring the extent of mental health symptoms. Each trial could be viewed as having a partially nested therapist design, with counsellors in the intervention but not the control arm. There was a single counsellor per patient. This meta-analysis gave a standardised mean difference (SMD) of -0.24 (95% CI -0.38 to -0.10). The primary meta-analysis assumed a common underlying treatment effect across trials (i.e. a fixed-effects meta-analysis model)

while a sensitivity analysis assumed the population treatment effects were normally distributed (i.e. a random-effects meta-analysis model). Neither made allowance for within-trial clustering due to counsellors or for between-arm heteroscedasticity. As four of the trials [34-37] reported the Beck Depression Inventory (BDI) [38], allowing a meta-analysis of the absolute mean differences, this subset will serve to illustrate the methods outlined below.

The BDI is one of the most widely used instruments for measuring the severity of depression. It is a 21-item self-report questionnaire, with total scores ranging from 0 to 63. Higher scores indicate more severe depressive symptoms. While a minimally important clinical difference for the BDI in this population has not been defined, a change of 4 to 5 points, corresponding to 0.5 standard deviations, is generally regarded to be minimally important. Although the trials all had partially nested designs, Friedli et al [35] and King et al [36] used a treatment manual, training or monitoring to standardise the delivery of counselling. Chilvers et al [34] and Simpson et al [37], instead, took a pragmatic approach. Patient eligibility was restricted to depression, or comorbid depression and anxiety, in Chilvers et al [34], King et al [36] and Simpson et al [37]. Friedli et al [35], in contrast, accepted a broad set of referrals. As such, this subset of trials also serves to illustrate meta-regression models for the variance parameters.

3. STATISTICAL MODELLING OF TWO-LEVEL NESTED TRIALS

First consider a cluster-randomised trial in which J clusters are randomly allocated to one of two treatments, with the only source of clustering in a fully nested design being recruitment-related. Suppose y_i is a continuous outcome for the i^{th} patient, where $i = 1, \ldots, N$, θ is the treatment effect, x_i and β are matrices signifying fixed patient or cluster level baseline covariates and their coefficients and K_i is an indicator variable for the intervention versus control. For simplicity of presentation let α_i equal $\alpha + \beta x_i$ where α is the constant. Using Goldstein's [39] notation, between-cluster variation can be represented by a random effect $u^{(2)}_{cluster(i)}$ with distribution $N[0, \sigma^2_u]$; $e^{(1)}_i$ is $N[0, \sigma^2_e]$, the patient level error term. A random-intercept model for the outcome for the i^{th} patient in the k^{th} treatment is therefore appropriate given by

$$y = \quad (1)$$

In this notation, the bracketed superscript refers to the level of the random effect and $\text{cluster}(i)$ in the subscript is the mapping of patients to clusters. Intra-cluster variability is measured by a single intraclass correlation coefficient ρ defined using a variance components model by $\rho = \sigma_u^2/(\sigma_u^2+\sigma_e^2)$.

Consider now any randomised trial in which care providers are allocated to patients within two treatments (k=0, 1) in a fully nested design. In the context of an individually randomised trial, Roberts and Roberts [6] suggest the following two-level heteroscedastic model

$$y_i = \beta_0 + \beta_1 \text{treat}_i + u^{(2)}_{\text{therapist}(i)0}(1-\text{treat}_i) + u^{(2)}_{\text{therapist}(i)1}\text{treat}_i + e^{(1)}_{i0}(1-\text{treat}_i) + e^{(1)}_{i1}\text{treat}_i \quad (2)$$

Model (2) would also be appropriate for a cluster randomised trial in which care providers are randomly allocated to two treatments, because one source of clustering is treatment provision and therefore treatment-related. In this parameterisation, $u^{(2)}_{\text{therapist}(i)0}$ and $u^{(2)}_{\text{therapist}(i)1}$ are random intercepts for the control and intervention arms respectively, distributed $N[0, \sigma^2_{u0}]$ and $N[0, \sigma^2_{u1}]$, with covariance zero, as they relate to independent samples. Note there are also separate patient level error terms for the control and intervention arms rather than just one across arms, given respectively by $e^{(1)}_{i0}$ and $e^{(1)}_{i1}$, and distributed $N[0, \sigma^2_{e0}]$ and $N[0, \sigma^2_{e1}]$, included to prevent bias in the estimation of σ^2_{u0} and σ^2_{u1} [6]. Separate intraclass correlation coefficients under a variance components model are then $\rho_0 = \sigma^2_{u0}/(\sigma^2_{u0}+\sigma^2_{e0})$ and $\rho_1 = \sigma^2_{u1}/(\sigma^2_{u1}+\sigma^2_{e1})$.

Where an individually-randomised trial has a partially-nested design, the random intercept for the control arm is constrained to equal zero so $u^{(2)}_{\text{therapist}(i)0}(1-\text{treat}_i)$ is dropped from the model giving

$$y_i = \beta_0 + \beta_1 \text{treat}_i + u^{(2)}_{\text{therapist}(i)1}\text{treat}_i + e^{(1)}_{i0}(1-\text{treat}_i) + e^{(1)}_{i1}\text{treat}_i \quad (3)$$

Each patient in the treatment arm without clustering is assumed to be a cluster of size one.

4. SUMMARY-DATA META-ANALYSIS METHODS

4.1 Fixed- and Random-Effects Meta-Analysis Models without Clustering

In the simplest meta-analysis model, an underlying treatment effect θ common to all H trials is assumed, such that $\theta = \theta_1 = \cdots = \theta_H$. The fixed-effects model [40] implies

$$\hat{\theta}_h = \theta + e_h, \quad h = 1, \ldots, H \qquad (4)$$

where $\hat{\theta}_h$ is the treatment effect observed in trial h, θ is the population value, and e_h are the sampling errors, with $e_h \sim N[0, \sigma^2_{\hat{\theta}_h}]$. Heterogeneity in the treatment effects across trials is ascribed to sampling error. The arguably more realistic random-effects model permits the population treatment effects to vary across trials, with $\theta_h = \theta + \varepsilon_h$ and $\theta_h \sim N[\theta, \tau^2_{\theta_h}]$, where $\tau^2_{\theta_h}$ is the between-trial variance and θ is now the mean of the population treatment effects. Thus [40]

$$\hat{\theta}_h = \theta + \varepsilon_h + e_h, \quad h = 1, \ldots, H \qquad (5)$$

and $\hat{\theta}_h \sim N[\theta, \sigma^2_{\hat{\theta}_h} + \tau^2_{\theta_h}]$. The total variance of $\hat{\theta}_h$ is therefore $T^2_{\hat{\theta}_h} = \sigma^2_{\hat{\theta}_h} + \tau^2_{\theta_h}$, the sum of the within and between trial variances. The random-effects model reduces to a fixed-effects meta-analysis model when $\tau^2_{\theta_h}$, the between trial variance, is zero.

The uniformly minimum-variance unbiased estimate of a pooled treatment effect θ is given by [41-42]

$$\hat{\theta}_w = \frac{\sum_{h=1}^{H} w_h \hat{\theta}_h}{\sum_{h=1}^{H} w_h} \qquad (6)$$

where $w_h = \frac{1}{T^2_{\hat{\theta}_h}}$ is the weight assigned to trial h under a random-effects meta-analysis model. Its standard error is given by

$$\sigma_{\hat{\theta}_w} = \sqrt{\frac{1}{\sum_{h=1}^{H} w_h}} \qquad (7)$$

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

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

It is usual for $\sigma^2_{\hat{\theta}_h}$ and $\tau^2_{\theta_h}$ to simply be replaced by their respective estimators $\hat{\sigma}^2_{\hat{\theta}_h}$ and $\hat{\tau}^2_{\theta_h}$, although Sidik and Jonkman [43] suggest an alternative approach that is robust to sampling errors in the estimated weights.

A commonly used estimator of $\tau^2_{\theta_h}$ is DerSimonian-Laird's (D-L) [44] methods of moments estimator

$$\hat{\tau}^2_{\theta_h} = \max\left[0, \frac{Q - (H-1)}{\sum_{h=1}^{H}\hat{\sigma}^{-2}_{\hat{\theta}_h} - \frac{\sum_{h=1}^{H}\left(\hat{\sigma}^{-2}_{\hat{\theta}_h}\right)^2}{\sum_{h=1}^{H}\hat{\sigma}^{-2}_{\hat{\theta}_h}}}\right] \qquad (9)$$

The Q-statistic is estimated by $\sum_{h=1}^{H}\hat{\sigma}^{-2}_{\hat{\theta}_h}\left(\hat{\theta}_h - \bar{\theta}_h\right)^2$, where $\bar{\theta}_h$ is the mean of $\hat{\theta}_h$. Variation in the precision of the trial estimates between trials is indexed by $\hat{I}^2 = \frac{\sum_{h=1}^{H}\hat{\sigma}^{-2}_{\hat{\theta}_h} - \frac{\sum_{h=1}^{H}\left(\hat{\sigma}^{-2}_{\hat{\theta}_h}\right)^2}{\sum_{h=1}^{H}\hat{\sigma}^{-2}_{\hat{\theta}_h}}}{\sum_{h=1}^{H}\hat{\sigma}^{-2}_{\hat{\theta}_h}}$.

In order to obtain the standard error of the absolute mean difference from each trial, $\hat{\sigma}^2_{\hat{\theta}_h}$, (used in calculating trial weights and the standard error of the pooled treatment effect), one needs to first derive the sampling distribution of the absolute mean difference. Where outcomes are statistically independent within and across arms, suppose μ_{1h} and μ_{0h} are the true mean outcomes in the intervention and control arm of trial h respectively. The population mean difference is then

$$\theta_{MD} = \mu_{1h} - \mu_{0h} \qquad (10)$$

The outcome of patient i in the k^{th} arm of the h^{th} study is denoted by y_{ikh}. Assuming the population variances are homogeneous ($\sigma^2_{1h} = \sigma^2_{0h} = \sigma^2_h$) and the sample means ($\bar{y}_{1h}$ and

$\bar{y}_{0h}$), variances (s_{1h}^2 and s_{0h}^2) and sizes (n_{1h} and n_{0h}) available, the trial estimate and its sampling distribution are given by [45]

(11)

where $\hat{\sigma}^2_{\hat{\theta}_{MD,h}} = s_h^2\left(\frac{1}{n_{1h}} + \frac{1}{n_{0h}}\right)$ and $s_h^2 = \frac{(n_{1h}-1)s_{1h}^2 + (n_{0h}-1)s_{0h}^2}{n_{1h}+n_{0h}-2}$

If the outcome variances are heterogeneous across arms (i.e. $\sigma_{1h}^2 \neq \sigma_{0h}^2$) with unknown ratio, the trial estimate $\hat{\theta}_{MD,h}$ is unaffected but its variance becomes

$$\hat{\sigma}^2_{\hat{\theta}_{MD,h}} = \frac{\sigma_{1h}^2}{n_{1h}} + \frac{\sigma_{0h}^2}{n_{0h}} \qquad (12)$$

The variances are replaced by s_{1h}^2 and s_{0h}^2 to give the estimator $\hat{\sigma}^2_{\hat{\theta}_{MD,h}}$, a scenario that is classically referred to as the Behrens-Fisher problem [46].

4.2 Sampling Distribution of the Summary Statistic for Two-Level Nested Designs

Suppose now that the outcome of patient i is nested within the j^{th} cluster of arm k and is denoted by y_{ijkh}. For the sake of generality, assume that model (2) applies. Then assume, for each of h trials, that a sample of J_{kh} clusters of size m_{kh} is assigned to each arm under a fully nested design. The trial estimate $\hat{\theta}_{MD,h} = \bar{y}_{1h} - \bar{y}_{0h}$ remains an unbiased estimator of $\theta_{MD,h}$ but the sample means are now given by

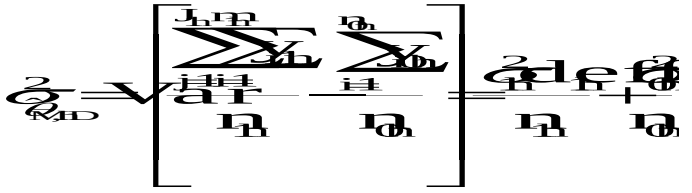
$$\bar{y}_{kh} = \frac{\sum_{j=1}^{J_{kh}}\sum_{i=1}^{m_{kh}} y_{ijkh}}{\sum_{j=1}^{J_{kh}} m_{kh}} = \frac{\sum_{j=1}^{J_{kh}}\sum_{i=1}^{m_{kh}} y_{ijkl}}{n_{kh}}, \text{ with sample variances } \hat{\sigma}^2_{\bar{y}_{kh}} = \frac{\sigma^2_{kh}\,\mathrm{deff}_{kh}}{n_{kh}} \qquad (13)$$

where the design effect $\mathrm{deff}_{kh} = 1 + (m_{kh}-1)\bar{\rho}_{kh}$, in the clustered arms when the cluster sizes are equal within each arm of each trial.

For this scenario, Kwong and Higgins [unpublished] gave the sampling distribution of $\hat{\theta}_{MD,h}$ as

(14)

where  . The sampling variance simplifies to

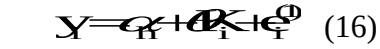


(15)

in the case of partial nesting and to  for cluster randomised trials where the only source of clustering is recruitment-related.

5. INDIVIDUAL-PATIENT-DATA META-ANALYSIS METHODS

Going back to Goldstein's [39] notation, where y_i denotes a continuous outcome for the i^{th} patient, a standard fixed-effects meta-analysis model [40, 47] is



(16)

where α_h represents the mean outcome in the control arm of trial h and θ the fixed treatment effect. It is commonly assumed that patient residuals $e_i^{(1)}$ are iid $N[0, \sigma_e^2]$, although relaxing this has been discussed [40, 47]. It is also possible to let the patient variance vary across arms, in which case the model becomes

(17)

with the $e_{ik}^{(1)}$ iid $N[0, \sigma_{ek}^2]$. This model can be extended to give the fixed-effects meta-analysis corresponding to a two-level heteroscedastic model by combining model (16) with that given by equation (2),

(18)

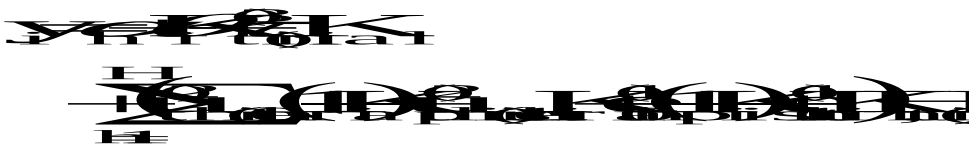
with the random effects iid $N[0, \sigma^2_{uk}]$. If all the trials are partially nested, $u^{(2)}_{therapist(i)0}$, can be omitted from the model, corresponding to equation (3).

A standard random-effects meta-analysis is one in which the trial effects are fixed but the treatment effect is permitted to vary randomly across trials [40, 47]. That is, the term $\tau^{(3)}_{trial(i)}K_i$ is added to model (16), where the $\tau^{(3)}_{trial(i)}$ are iid $N[0, \tau^2]$ and the random effects are mutually independent. The random-effects meta-analysis corresponding to a two-level heteroscedastic model for the trials is given by

 (19)

As before, $u^{(2)}_{therapist(i)0}$ is constrained to zero, and the term omitted from the model, if all trials are partially nested.

Models (18) and (19) constrain the therapist variance to be equal across trials for each treatment. An alternative would be a saturated model in which all trials are allowed to have their own therapist variance. Suppose $H_{h,trial(i)}$ is an indicator variable equal to 1 when $trial(i)=h$ and 0 otherwise, the saturated model can be defined as follows:

 (20)

With $4H$ variance parameters in a meta-analysis of fully-nested trials and $3H$ variance parameters in a meta-analysis of partially-nested trials, Model (20) is likely to be difficult to fit. It was not possible in our motivating example. One option is to add constraints to the saturated model that can be motivated by the characteristics of the trials, a possibility we now consider.

6. META-REGRESSION MODELS USING INDIVIDUAL-PATIENT-DATA

Meta-regression models have been described that allow the pooled treatment effect to vary according to trial characteristics [47-49], such as whether the trial intervention was manualised or the trial quality. These models explore explanations for between-trial variation and require large numbers of trials. Incorporation of a categorical trial-level covariate into model (19) gives

 (21)

where β is a fixed treatment-by-covariate interaction effect and x_h is an indicator variable for the fixed trial-level covariate.

Further covariates could be added. Where data are available on therapist-level characteristics such as training or experience, one might be interested in exploring whether the treatment effect varies according to these. Here, the covariate varies within trials, but is the same for every patient seen by a therapist. As the number of therapists per trial is usually small, it may only begin to be feasible to address such questions in a meta-regression. As with other IPD meta-regressions, patient-level covariates, such as severity, can also be investigated [47]. In this case, the covariate varies between patients within therapists and trials.

Up to this point, the meta-regressions considered are of fixed effects, and in particular of the treatment effect. Meta-regressions of random parameters may also be of interest. A complex random structure may be realistic if the trial designs vary. Under these circumstances, there is reason to expect between-trial variation in therapist or patient level random effects even if there is insufficient statistical power available to detect it. It is realistic to suppose that patient and therapist level variances are affected by standardising patient or therapist characteristics and behaviour via the use of selection criteria and therapist training, certification, monitoring and supervision. If the trial designs are comparable in all other respects, a categorical trial-level covariate can be incorporated for the therapist random intercept in model (19). T_i is an indicator variable that is equal to 1 if therapist characteristics or behaviour are standardised and 0 otherwise, for example, as follows,

(22)

where the four $u^{(2)}_{therapist(t)kt}$ are random intercepts for the control and intervention arms (k=0,1) in the unstandardised and standardised trials (t=0,1) respectively, distributed $N[0, \sigma^2_{ukt}]$, with covariance zero, as they relate to independent samples.

This might be considered if some of the trials used treatment manuals, while others did not, or if therapists were selected for their expertise, given training, accreditation, monitoring or supervision in some trials but not others. It is assumed that these design features do not have a simultaneous effect at the patient level in Model (22), as this leads to the saturated model (20) in our motivating example. One could instead incorporate a categorical trial-level covariate for the patient-level residual error. For example, P_i is 1 for trials where patient characteristics are standardised and 0 otherwise,

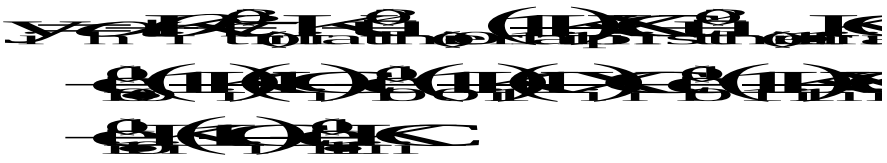
(23)

where the four $e^{(1)}_{ikp}$ are patient residuals for the control and intervention arms (k=0,1) in the unstandardised and standardised trials (p=0,1) respectively, distributed $N[0, \sigma^2_{ekp}]$, with covariance zero, as they again relate to independent samples. This might be considered if trials adopt a mix of explanatory and pragmatic approaches to patient eligibility. Models (22) and (23) may be considered parsimonious or constrained versions of the saturated model (20). The potential complexity increases with the variability in the trial designs. If the number of trials is small, as we have seen, there may be a trade-off between a realistic model for the random effects and computational feasibility. In theory, these models could be extended to include therapist- and patient-level predictors of the random effects.

As an aside, Model (20) can also be simplified to allow inclusion of fully and partially nested trials and inclusion of trials with and without clustering effects. In the case of a mixture of fully and partially nested designs, where X_i is an indicator variable equal to 1 when the trial has a fully nested design and 0 if it is partially nested,

(24)

Here, the residual error in the control arm is allowed to differ across trial designs, ensuring the therapist ICC in the control arm is based on the subset of trials with fully nested designs. As before, it is assumed that the therapist ICC in the control arm is homogeneous for all fully nested trials. If the independence assumption is reasonable in some of the trials, Model (24) can be extended, with C_i an indicator variable equal to 1 if a trial has any clustering effects and 0 otherwise, to give



(25)

Each random intercept applies only to the clustered arms. The residual error again varies by trial design. For non-clustered trials, it is $\sigma^2_{w0}(1-K_i)(1-C_i)$ in the control arm and $\sigma^2_{w1}K_i(1-C_i)$ in the intervention arm. The latter term can be omitted if the patient-level variance is assumed to be homogeneous across arms.

An, albeit rather contrived, example in which fully-, partially-nested and non-clustered trials might be pooled is a comparison between counselling and cognitive-behavioural therapy where both have web-based and face-to-face versions. Some trials might compare web-based versions, thereby incorporating no therapist involvement, and so be non-clustered. Others might compare face-to-face versions to web-based versions and be partially- nested. Others might compare the face-to-face versions and be fully-nested. Another situation in which one might be justified in considering Model (25) is when the number of therapists cannot be identified in one or more of the trials. In this case they may be included as non-clustered.

7. APPLICATION TO THE MOTIVATING EXAMPLE

Short-term outcomes relating to the Beck Depression Inventory (BDI) were available for 460 patients from four [34-37] of the counselling in primary care trials. Of these, 224 (49%) were allocated counselling with one of 39 counsellors. Overall, the cluster sizes ranged from 1 to

33, with a median of 3 and an IQR of 1 to 8. Data were available for 5 or more patients for 18 of the counsellors. Table 1 gives descriptive statistics for the four included trials. It can be seen that the trials with the largest treatment effects also had the smallest counsellor ICCs. The ANOVA estimates of the counsellor ICC are negative for two of the four trials. This is possible because 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 since a between-cluster variance cannot be negative. It is the design effect rather than the ICC that cannot be negative in ANOVA estimation. If clusters are of size two, the range of the ICC is ± 1, but as the cluster size increases the minimum approaches zero. One ICC across trials was initially assumed for the counselling arm.

[Insert Table 1 about here]

7.1 Summary-Data versus Individual-Patient-Data Meta-Analyses

To reflect a common lack of knowledge about the cluster size distribution, equal cluster sizes within trials were assumed for all summary-data meta-analyses. A pooled ICC of 0.033 was used, based on a weighted average of the trial-specific ICCs [10], regardless of the model. IPD models were implemented in MLwiN using RIGLS, due to its flexibility in modelling random effects. RIGLS is comparable to REML [39] implemented in mixed in Stata Version 13. The preceding command xtmixed was updated in Version 11 to permit inclusion of one covariate for the patient level error. The mixed command uses the same syntax but seems to be faster, with a more stable algorithm. Details of the programming for both packages are given as supporting web materials.

Tables 2 and 3 summarise, respectively, the summary-data and IPD estimates and standard errors for the fixed- and random-effects meta-analyses, progressively relaxing independence and common variance assumptions within the trials. As all of the trials have partially nested designs, the Level 2 variance, where it applies, is heterogeneous in all analyses. The common variance assumptions therefore relate only to the Level 1 variance. As can be seen, the pooled mean difference and its standard error for a usual summary-data fixed-effects analysis are -2.43 and 0.89 (95% CI -4.17 to -0.69), indicating that counselling reduces short term symptoms of depression by an average of 2.4 points and that this reduction is statistically significant at the 5% level. A mean difference of 2.5 points corresponds to a standardised effect size of about 0.25. According to Cohen [51] this represents a small effect. Based on

results with similar effects, the authors of the Cochrane review concluded that "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" [52]. The equivalent IPD estimate and its standard error are -2.47 and 0.90 with the two-sided 95% CI -4.23 to -0.71. The similarity of these results implies that bias and sampling error in the summary-data within-trial variance estimates is not important here. The pooled mean difference and its standard error in the analogous summary-data random-effects analysis are -2.50 and 1.40 (95% CI -5.24 to 0.24). The increase in standard error arises from between-trial heterogeneity in the mean differences. The reduction in BDI is no longer statistically significant. If an IPD approach had been used, the estimate and its standard error would be -2.47 and 1.42 (95% CI -5.25 to 0.31). The slight disparity in standard errors is explained by that of the between-trial variance estimates, which is in turn due to bias arising from sampling error or heterogeneity in the within-trial variances. Even so, the evidence in favour of counselling in primary care is less clear if between-trial heterogeneity is taken into consideration.

[Insert Tables 2 and 3 about here]

The impact of between-arm heteroscedasticity and within-trial clustering is minimal if pooled treatment effects and their standard errors are compared across random-effects summary-data or IPD models (see Figure 1 below). The effect is a little more pronounced for both summary-data and IPD fixed-effects models, however. The disparity between the summary-data and IPD results enlarges as the model becomes more realistic. It is of note that the DerSimonian-Laird (D-L) and IPD between-trial variance estimates differ (see Table 3), with both estimates being larger, assuming independence, where patient-level variances are allowed to differ between arms. The IPD estimate, in contrast to the D-L estimate, is not only smaller for both clustered models but also smaller for the clustered model where patient-level variances are allowed to differ between arms. IPD estimates of the counsellor ICC are larger than the summary-data estimate of 0.033, varying from model to model. These differences arise, in part, because the variances are estimated simultaneously in an IPD model, making appropriate allowance for all other effects in the model. In this example, the results continue to be dominated by between-trial heterogeneity in the treatment effects. The most realistic IPD pooled mean difference and standard error are -2.51 and 1.45 (95% CI -5.35 to 0.33). The summary-data equivalent is -2.53 and 1.43 (95% CI -5.33 to 0.27). Both are very similar.

In the IPD case, the confidence interval is marginally wider than the standard random effects one. The conclusion remains unchanged.

[Insert Figure 1 about here]

7.2 Sensitivity of the Summary-Data Approach to the Choice of Population ICC

The sensitivity of the mean difference and its standard error to the choice of population ICC was explored for ICCs between zero and one. The trial estimates are unaffected as the ICC increases but the pooled estimates become slightly more extreme. This is because King et al [36] has more weight as the ICC increases, in part due to its mean cluster size. This effect is slightly more pronounced for the fixed-effects estimate. The slope of the pooled standard error, when plotted against the population ICC, is not steep, indicating the results are not sensitive to the ICC in the anticipated range (i.e. for ICCs between zero and 0.20). The D-L estimate of $\tau^2_{\theta_h}$ decreases as the ICC increases, implying heterogeneity in mean differences across trials contributes to, rather than simply explaining, heterogeneity between counsellors.

7.3 Meta-Regression of the Random Effects

Table 4 gives results of two meta-regression models, one for the therapist random effect (Model 22) and the other for the patient residual (Model 23). Both explore trial-level sources of heterogeneity in the counsellor ICC, the first treatment standardisation (yes, no) and the second patient eligibility (mixed diagnosis, depression). There were insufficient trials available to fit random-effects meta-regression models in this instance so the results are compared to model (18). As all the trials have partially nested designs, the random intercept for the control arm, $u^{(2)}_{\text{therapist}(t)0}$, is omitted from all models.

[Insert Table 4 about here]

A reduction of 8.7 was seen in the log likelihood by including separate residual terms for trials with mixed and depression patient referrals. The pooled treatment effect reduced very slightly, as did its standard error. The counsellor ICC was higher when patients were more homogeneous as the patient residual was smaller relative to the counsellor variance. If distinct therapist-level terms were included for trials standardising counselling and those that did not, the log likelihood reduced by 12.6. The pooled treatment effect increased

appreciably, reflecting an association between the trial estimate and counsellor ICC. That is, the trials with the largest estimates (i.e. Friedli [35] and King [36]) also had the smallest counsellor ICCs, so carried more weight in the meta-regression analysis. The standard error was similar to that for Model (18). Since the pooled counsellor ICC is negative for trials standardising counselling, a different parameterisation of the model was used, including a covariance term rather than an explicitly negative estimate, to allow the model to converge. A covariance, in contrast to a variance, can be negative. Including the covariance between the therapist level random effects in place of the negative variance therefore indirectly enabled a negative variance to be estimated within a variance components model. The counsellor ICC was lower when counselling was standardised as the counsellor variance was smaller relative to the patient residual. This corresponds to the ANOVA estimates of the ICC in Table 1. The standard errors for the variance estimates are large due to the number of trials and counsellors. It was not computationally possible to simultaneously allow for heterogeneity from both sources (i.e. Models 22 and 23 combined) or to fit the model of choice (i.e. a random-effects meta-regression). The potential to do so when the number of trials available is larger is clear however. The facility to disentangle the predictors of the components of an ICC is also attractive as the predictors may differ between the components.

8. DISCUSSION

While potentially important, treatment-related clustering effects in individually-randomised psychotherapy trials have rarely been taken into account in trial reports and do not appear to have been considered in meta-analyses [10]. Fitting fixed- and random-effects meta-analysis models to trials of counselling in primary care, adopting summary-data and IPD approaches and allowing for these effects, had minimal impact on the pooled estimate and its standard error. This is not surprising for two reasons. Firstly, the cluster sizes were small in the example so the design effect was also. Secondly, assuming a common ICC across trials in the counselling arm meant that the contribution of each trial to the pooled treatment estimate remained essentially the same, despite some variability in the mean cluster size. Although hardly noticeable in the example, the impact was instead on the precision of the pooled treatment effect. Nevertheless, as we have seen in Tables 2 and 3, failure to take account of therapist variation will give an overly precise pooled estimate in a fixed effects meta-analysis because the effect of failing to include a therapist random effect in the analysis of a single trial generally results in the variance of the treatment effect being underestimated. The picture

is more complex for a random-effects meta-analysis. If the variance of single trials is underestimated, the between trial variance may be overestimated, as was seen in Tables 2 and 3. The combined effect of a reduction in the variance between trials and an increased variance of each trial can result in either a reduction or an increase in the standard error of the random-effects pooled estimate. Whilst in our example this estimate had a marginally larger variance when therapist clustering had been taken in to account, a different set of cluster sizes or trial variances could have led to a reduction.

By contrast, an appreciable impact of treatment-related clustering was observed on the pooled treatment effect in the meta-regression models. Here, between-trial heterogeneity in the counsellor ICC had a greater impact on the weight given to particular trials and in so doing affected the pooled estimate and its standard error. Collection of the IPD is made attractive by the potential of meta-regression analyses for exploring trial-, therapist- and patient-level predictors of the treatment effect and of the random effects. Increased sample sizes open up opportunities not usually present at a trial-level but computational problems may still arise largely due to the presence of negative estimates. Allowing the ICCs to vary by trial as well as by treatment arm is particularly likely to lead to problems, as many of the trial-level ICC estimates involve very small numbers of clusters. The middle road suggested here is one way of circumventing these problems while maintaining a more realistic model.

An advantage of the proposed methods is their generality. A two-level heteroscedastic model relaxes common variance and independence assumptions, being appropriate for all fully nested designs. It simplifies to the models recommended for unequal patient-level variances across arms and for partially-nested, cluster-randomised and non-clustered designs. In each of these special cases, additional assumptions may be made so constraints can be added to the model at the trial level. It is possible to envisage scenarios where one might want to allow the ICC to vary by treatment arm in a cluster randomised trial. Here, the source of clustering is traditionally conceptualised as recruitment-related. For example, if at baseline GP practices, rather than patients, are randomised to treatments, patients within a GP practice are likely to be more similar to one another than to other patients in the trial. As a consequence, clustering arises from use of a two-stage or clustered sample in a cluster-randomised trial but not in an individually-randomised trial. Such clustering is expected to be maintained at follow-up.

The unit of randomisation may not be the only source of clustering in a cluster randomised trial however, particularly where the intervention is directed at the cluster (e.g. GP practice) rather than at the patient-level. If there is also treatment-related clustering then, as long as the unit of randomisation and the clusters relating to treatment are the same (e.g. GPs are the unit of randomisation and the care providers), a two-level heteroscedastic model, outlined here for a fully-nested therapist design, may be appropriate in a cluster-randomised trial even if the treatment-related clustering is restricted to one or more arms in the trial. Consider a trial in which groups of patients are cluster-randomised to intervention or control, where the intervention is some kind of group therapy and the control is no therapy as an example. Clustering related to recruitment would still apply in the control arm, where in an individually-randomised trial it may be constrained to zero, but you might not expect the clustering effect to be equal in both arms as you might in a traditional cluster-randomised trial. Where there is interest in comparing group therapy to no therapy, one might want to consider pooling trials using an individually- and a cluster-randomised design with meta-regression models similar in principle to those described here. The general principle we have adopted is that the cluster and patient-level components of an ICC should be allowed to differ by trial design, at a minimum.

In the motivating example there was also the potential for clustering by the GP. GP care was generally a co-intervention delivered by the same sample of GPs. As such, GPs were crossed with treatment arms. As they were not blinded to whether patients were allocated counselling or no counselling, an interaction between GPs and treatment arm is plausible. Information on GP involvement in the motivating example was very limited however. GP identifiers were not recorded for the majority of the trials so it was not straightforward to include GPs in IPD analyses nor was it often possible for researchers to report the level of between-GP variability in the treatment effect. The number of GPs treating trial patients was also often unavailable so there was very limited information on cluster size distributions. As such, while a literature is starting to develop on the statistical implications of multiple therapist-per-patient designs [53], it is likely to be generally the case that details of multiple therapists treating particular participants are unavailable in this setting. This is likely to be true of multiple therapists of the same type (e.g. if more than one counsellor had treated patients) or of different types (e.g. in the case of a counsellor and a GP, as was the case here), even though both multiple therapist-per-patient trial designs are common in psychotherapy [10]. That is, trials in which the relationship between therapists and patients can be described as “multiple-membership”

or "cross-classified" [5]. Extensions are needed to the methods proposed here for these more complex data structures, as well as for crossed designs and trials with further levels (e.g. centres) or repeated measurements over time.

An important consideration when implementing the summary data methods proposed here is the feasibility of obtaining, by trial arm, the ICC and average cluster size when researchers have made no allowance for clustering by care providers. To our knowledge, ICC estimates are currently only very rarely reported in the principal reports of psychotherapy trials [e.g. 56]. Subsequent papers may be published focusing on therapist effects, such as a series of papers relating to the NIMH Treatment for Depression Collaborative Research Program trial [55-58], or for the purpose of generating a database of therapist effects [59-60]. The number of therapists involved in a psychotherapy trial is commonly reported though and tends to be no greater than ten per arm. It is therefore likely to be possible to calculate average cluster sizes. The distribution of the cluster sizes may however be skewed and highly variable, with only a few therapists treating the majority of participants, as was the case in Chilvers et al [34] and King et al [36]. As this is not likely to be clear from the principal paper, no allowance was made for it in the methods described here. More generally, variability in cluster sizes within trials is likely to be common, and while it is difficult to make appropriate allowance for this if the cluster size distribution is unknown, the assumption of equal cluster sizes is a limitation of our methods. For these reasons, the IPD approach is preferred, but this assumes researchers are able to link clusters to participants.

From experience of collecting the therapist data for this meta-analysis, it is likely that cluster identifiers are collected in the paper records of psychotherapy trials and it is common for them to be somewhere in the electronic dataset. Although time consuming, it was possible to get hold of IPD for all the trials of counselling in primary care. Contact started with the lead author of the Cochrane review and progressed to the lead author (and statistician where appropriate) for each trial. In two of seven trials, data was re-entered from the paper case report forms. This was the entire dataset for one but for the other it was just the counsellor identifiers. Every trial recorded the counsellor who provided treatment. The age of the trial is likely to be a factor in how accessible data is more generally. Establishing a collaborative group and making use of personal contacts both helped to facilitate permissions to use data. In other meta-analyses, it is possible that only the summary-data will be available in one or more eligible trials. Where this is the case, assumptions could be made about the size of the

clustering effect and sensitivity analyses carried out using a summary data approach. Further work is needed to extend formal methods for pooling a mixture of IPD and summary-data in this context [61-62]. As psychotherapy researchers have been interested in therapist effects for a number of decades, the availability of ICCs, cluster sizes and cluster identifiers is likely to be greater in this context than in nested trials of occupational therapy, surgery or physiotherapy for example. Where this data is unavailable, as may often be the case in some areas, assumptions could be made and sensitivity analyses carried out.

The focus of this paper has been on meta-analyses of absolute mean differences in the context of a three-level model (patients are nested within clusters, nested within trials) appropriate for trials with nested designs. In the situation where a normally-distributed outcome is measured with different questionnaires or scales across trials, say depression on the BDI, HADS-D and PHQ-9, a standardised mean difference would be the appropriate measure of treatment effect. The methods described in this paper do extend but there are a number of added complications that must be taken into account relating to small-sample bias in the treatment effect estimate, estimates having a non-central t-distribution and dependence of the sampling variance on the population parameter. A separate paper is in preparation focusing on the specific issues with pooling standardised mean differences in this context. It is important that specific issues arising in the context of standardised mean differences, odds ratios, relative risks, and hazard ratios are fully considered, since allowing for treatment-related clustering is more complex for these summary statistics. (It would be easier for estimates of risk difference as one would simply multiply the standard error by a design effect term here). One of these specific issues is that population-averaged or marginal estimates will be required if a summary-data approach is adopted, rather than cluster-specific or conditional ones [63]. The 95% CIs presented here all used the z-statistic from equation (8). Where small samples of trials or therapists are pooled, it will be more appropriate to use a t-statistic [43, 64]. The degrees of freedom relating to this statistic for a random-effects meta-analysis are based on the number of trials. In the case of a fixed-effects meta-analysis, they are more complex, being based on a Satterthwaite approximation [65] required for meta-analysing standardised mean differences [10].

In conclusion, specific guidance is needed in the Cochrane Handbook [27] on methods for handling treatment-related clustering associated with care providers in either individually- or cluster-randomised trials. We have shown that while the issues may have similarities to those

for standard cluster-randomised trials, the methods themselves need to be more general. While we have focused on implications for precision, this guidance should consider the implications for internal and external validity of pooled treatment effect estimates [10], as well as those for precision, as these affect interpretation and the validity issues are just as important.

REFERENCES

1. Donner A, Klar N. Design and Analysis of Cluster Randomization Trials in Health Research. Arnold: London, UK, 2000.
2. Giraudeau B, Ravaud P, Donner A. Sample size calculation for cluster randomized cross-over trials. Statistics in Medicine 2008; **27**: 5578-5585.
3. Rietbergen C, Moerbeek M. The design of cluster randomized crossover trials. Journal of Educational and Behavioral Statistics 2011; **36**: 472-490.
4. Turner RM, White IR, Croudace T, for the PIP Study Group. Analysis of cluster randomized cross-over trial data: A comparison of methods. Statistics in Medicine 2007; **26**: 274-289
5. Walwyn R, Roberts C. Therapist variation within randomised trials of psychotherapy: Implications for precision, internal and external validity. Statistical Methods in Medical Research 2010; **19**: 291–315.
6. Roberts C, Roberts SA. Design and analysis of clinical trials with clustering effects due to treatment. Clinical Trials 2005; **2**: 152–162.
7. Lee KJ, Thompson SG. Clustering by health professional in individually randomised trials. British Medical Journal 2005; **330**(7483): 142–144.
8. Roberts C. The implications of variation in outcome between health professionals for the design and analysis of randomized controlled trials. Statistics in Medicine 1999; **18**: 2605–2615.
9. Boutron I, Moher D, Altman DG, Schulz KF, Ravaud P, for the CONSORT Group. Extending the CONSORT statement to randomized trials of non-pharmacological treatment: explanation and elaboration. Annals of Internal Medicine 2008; **148**: 295–309.
10. Walwyn R. Therapist Variation within Meta-Analyses of Psychotherapy Trials. PhD Thesis. University of Manchester, UK, 2010.
11. Deeks JJ, Altman DG, Bradburn MJ. Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. In Systematic Reviews in Health Care: Meta-analysis in context, Egger M, Davey-Smith G, Altman DG (eds). BMJ Books: London, 2001.
12. Glass GV. Primary, secondary and meta-analysis of research. Educational Researcher 1976; **5**: 3-8.

13. Jones AP, Riley RD, Williamson PR, Whitehead A. Meta-analysis of individual patient data versus aggregate data from longitudinal clinical trials. Clinical Trials 2009; **6**: 16-27.
14. Morris SB. Effect size estimation from pretest-posttest-control designs with heterogeneous variances. 20th Annual Conference of the Society for Industrial and Organizational Psychology, 2005.
15. Morris SB. Estimating effect sizes from pretrest-posttest-control group designs. Organizational Research Methods 2008; **11**: 364-386.
16. Morris SB, DeShon RP. Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. Psychological Methods 2002; **7**: 105-125.
17. Curtin F, Altman DG, Elbourne D. Meta-analysis combining parallel and cross-over clinical trials. I: Continuous outcomes. Statistics in Medicine 2002; **21**: 2131-2144.
18. Curtin F, Elbourne D, Altman DG. Meta-analysis combining parallel and cross-over clinical trials. II: Binary outcomes. Statistics in Medicine 2002; **21**: 2145-2159.
19. Curtin F, Elbourne D, Altman DG. Meta-analysis combining parallel and cross-over clinical trials. III: The issue of carry-over. Statistics in Medicine 2002; **21**: 2161-2173.
20. Elbourne DR, Altman DG, Higgins JPT, Curtin F, Worthington HV, Vail A. Meta-analyses involving cross-over trials: Methodological issues. International Journal of Epidemiology 2002; **31**: 140-149.
21. Donner A, Klar N. Issues in the meta-analysis of cluster randomized trials. Statistics in Medicine 2002; **21**: 2971-2980.
22. Donner A, Piaggio G, Villar J. Statistical methods for the meta-analysis of cluster randomization trials. Statistical Methods in Medical Research 2001; **10**: 325-338.
23. Hedges LV. Effect sizes in cluster randomized designs. Journal of Educational and Behavioral Statistics 2007; **32**: 341-370.
24. Hedges LV. Effect sizes in nested designs. In The Handbook of Research Synthesis and Meta-Analysis (2nd Ed), Cooper H, Hedges LV, Valentine JC (eds). Russell Sage Foundation: New York, 2009.
25. White IR, Thomas J. Standardized mean differences in individually-randomized and cluster-randomized trials, with applications to meta-analysis. Clinical Trials 2005; **2**: 141-151.
26. Zou G. One relative risk versus two odds ratios: implications for meta-analyses involving paired and unpaired binary data. Clinical Trials 2007; **4**: 25-31.
27. Higgins JPT, Green S (eds). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, March 2011. Available from www.cochrane-handbook.org (accessed 20/12/2013).
28. Viechtbauer W. Bias and efficiency of meta-analytic variance estimators in the random-effects model. Journal of Educational and Behavioral Statistics 2005; **30**: 261-293.

29. London School of Economics Centre for Economic Performance Mental Health Policy Group. The Depression Report: A new deal for depression and anxiety disorders, 2006. Available from http://eprints.lse.ac.uk/archive/00000818 (accessed 20/12/2013).
30. Mellor-Clark J, Simms-Ellis R, Burton M. National Survey of Counsellors in Primary Care: Evidence for growing professionalisation? Royal College of General Practitioners: London, 2001.
31. Bond T. The nature and role of counselling in primary care. In Counselling in Primary Care, Keithley J, Bond T, Marsh G (eds). Oxford University Press: Oxford, 2002.
32. Rowland N. Counselling and counselling skills. In Counselling in General Practice, Sheldon M (ed). Royal College of General Practitioners: London, 1992.
33. Bower P, Rowland N. Effectiveness and cost effectiveness of counselling in primary care. Cochrane Database of Systematic Reviews 2006; Issue 3. Art. No.: CD001025. DOI: 10.1002/14651858.CD001025.pub2.
34. Chilvers C, Dewey M, Fielding K, Gretton V, Miller P, Palmer B, Weller D, Churchill R, Williams I, Bedi N, Duggan C, Lee A, Harrison G. 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): 772-775.
35. Friedli K, King MB, Lloyd M, Horder J. Randomised controlled assessment of non-directive psychotherapy versus routine general-practitioner care. Lancet 1997; **350**(9092): 1662-1665.
36. King M, Sibbald B, Ward E, Bower P, Lloyd M, Gabbay M, Byford S. 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): 1-83.
37. Simpson S, Corney R, Fitzgerald P, Beecham J. A randomised controlled trial to evaluate the effectiveness and cost-effectiveness of counselling patients with chronic depression. Health Technology Assessment 2000; **4**(36).
38. Beck AT, Ward C, Mendelson M, Erbaugh J. An inventory for measuring depression. Archives of General Psychiatry 1961; **6**: 561-571.
39. Goldstein H. Multilevel Statistical Models (3rd Edition). Arnold: London, 2003.
40. Whitehead A. Meta-Analysis of Controlled Clinical Trials. Wiley: New York, 2002.
41. Birge RT. The calculation of errors by the method of least squares. Phys Rev 1932; **16**: 1-32.
42. Cochran WG. Problems arising in the analysis of a series of similar experiments. Journal of the Royal Statistical Society 1937; **4 (Supplement)**: 102-118.
43. Sidik K, Jonkman JN. Robust variance estimation for random effects meta-analysis. Computational Statistics & Data Analysis 2006; **50**:3681-3701.
44. DerSimonian R, Laird NM. Meta-analysis in clinical trials. Controlled Clinical Trials 1986; **7**: 177-188.

45. Borenstein M. Effect sizes for continuous data. In The Handbook of Research Synthesis and Meta-Analysis (2nd Ed), Cooper H, Hedges LV, Valentine JC (eds). Russell Sage Foundation: New York, 2009.
46. Kim S-H, Cohen AS. On the Behrens-Fisher problem: A review. Journal of Educational and Behavioral Statistics 1998; **23**(4): 356-377.
47. Higgins JPT, Whitehead A, Turner RM, Omar RZ, Thompson SG. Meta-analysis of continuous outcome data from individual patients. Statistics in Medicine 2001; **20**(15): 2219-2241.
48. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: A comparison of methods. Statistics in Medicine 1999; **18**(20): 2693-2708.
49. Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? Statistics in Medicine 2002; **21**(11): 1559-1573.
50. Wang CS, Yandell BS, Rutledge JJ. The dilemma of negative analysis of variance estimators of intraclass correlation. Theoretical and Applied Genetics 1992; **85**: 79-88.
51. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Academic Press: New York, 1977.
52. Bower P, Rowland N, Hardy R. The clinical effectiveness of counselling in primary care: a systematic review and meta-analysis. Psychological Medicine 2003; **33**: 203-215.
53. Roberts C, Walwyn R. Design and analysis of non-pharmacological treatment trials with multiple therapists per patient. Statistics in Medicine 2013; **32**: 81–98.
54. Goodyer I, Dubicka B, Wilkinson P, Kelvin R, Roberts C, Byford S, Breen S, Ford C, Barrett B, Leech A, Rothwell J, White L, Harrington R. Selective serotonin reuptake inhibitors (SSRIs) and routine specialist care with and without cognitive behaviour therapy in adolescents with major depression: randomised controlled trial. BMJ 2007; **335**:142-149.
55. Crits-Christoph P, Gallop R. Therapist effects in the National Institute of Mental Health Treatment of Depression Collaborative Research Program and other psychotherapy studies. Psychotherapy Research 2006; **16**(2):178-181.
56. Elkin I, Falconnier L, Martinovich Z, Mahoney C. Rejoinder to commentaries by Stephen Soldz and Paul Crits-Christoph on therapist effects. Psychotherapy Research 2006; **16**(2):182-183.
57. Elkin I, Falconnier L, Martinovich Z, Mahoney C. Therapist effects in the National Institute of Mental Health Treatment of Depression Collaborative Research Program. Psychotherapy Research 2006; **16**(2):144-160.
58. Kim D-M, Wampold BE, Bolt DM. Therapist effects in psychotherapy: A random-effects modeling of the National Institute of Mental Health Treatment of Depression Collaborative Research Program data. Psychotherapy Research 2006; **16**(2):161-172.
59. Baldwin SA, Murray DM, Shadish WR, Pals SL, Holland J, Abramowitz JS, Andersson G, Atkins DC, Carlbring P, Carroll KM, Christensen A, Eddington KM, Ehlers A, Feaster DJ, Keijsers GPJ, Koch E, Kuyken W, Lange A, Lincoln T, Stephens RS, Taylor S, Trepka C, Watson J. Intraclass

correlation associated with therapists: Estimates and applications in planning psychotherapy research. Cognitive Behaviour Therapy 2011; **40**(1)

60. Crits-Christoph P, Baranackie K, Kurcias JS, Beck AT, Carroll KM, Perry K, Luborsky L, McLellan AT, Woody GE, Thompson L, Gallagher D, Zitrin C. Meta-analysis of therapist effects in psychotherapy outcome studies. Psychotherapy Research 1991; **1**(2): 81-91.
61. Riley RD, Lambert PC, Staessen JA, Wang J, Gueyffier F, Thijs L, Boutitie F. Meta-analysis of continuous outcomes combining individual patient data and aggregate data. Statistics in Medicine 2008; **27**: 1870-1893.
62. Sutton AJ, Kendrick D, Coupland CAC. Meta-analysis of individual- and aggregate-level data. Statistics in Medicine 2008; **27**: 651-669.
63. Bohning D, Malzahn U, Dietz E, Schlattmann P, Viwatwongkasem C, Biggeri A. Some general points in estimating heterogeneity variance with the DerSimonian-Laird estimator. Biostatistics 2002; **3**(4): 445-457.
64. Rosner B. A generalization of the paired t-test. Applied Statistics 1982; **31**: 9-13.
65. Satterthwaite FE. An approximate distribution of estimates of variance components. Biometrics Bulletin 1946; **2**(6): 110-114.

Table 1 Descriptive Statistics for the Short-Term Beck Depression Inventory for Counselling vs. Control

Trial	Counselling			No Counselling			Mean Cluster Size in Counselling Arm	Counsellor ICC (ANOVA Estimate)
	N	Mean	(SD)	N	Mean	(SD)		
Chilvers 2001	39	15.2	(11.6)	44	14.8	(10.1)	2.79	0.290
Friedli 1997	59	11.7	(7.7)	51	15.6	(10.5)	14.75	-0.023
King 2000	62	11.5	(7.7)	62	17.2	(11.9)	4.23	-0.140
Simpson 2000	82	16.0	(9.3)	79	16.0	(8.1)	8.88	0.045

Note: SD = standard deviation; CI = confidence interval; ICC = intraclass correlation coefficient; A weighted average of the four ICCs gave a pooled ICC of 0.033 (see [10])

Table 2 Fixed- and Random-Effects Summary-Data Meta-Analyses of the Absolute Mean Difference in BDI between Counselling and Control where All Trials have Partially Nested Designs

Fixed-Effects Meta-Analysis	**Assuming Independence**				**Allowing for Within-Trial Clustering**			
Level 1 Variance	**Equal**		**Unequal**		**Equal**		**Unequal**	
	(Model 4 with Variance 11)		**(Model 4 with Variance 12)**		**(Model 4)**		**(Model 4 with Variance 15)**	
Trial	**% Weights**	**Mean Difference (Standard Error)**	**% Weights**	**Mean Difference (Standard Error)**	**% Weights**	**Mean Difference (Standard Error)**	**% Weights**	**Mean Difference (Standard Error)**
Chilvers 2001	13.9	0.4 (2.38)	13.7	0.4 (2.40)	15.1	0.4 (2.41)	14.6	0.4 (2.44)
Friedli 1997	25.7	-3.9 (1.75)	24.8	-3.9 (1.79)	23.8	-3.9 (1.92)	23.9	-3.9 (1.91)
King 2000	22.9	-5.8 (1.85)	24.3	-5.8 (1.80)	24.2	-5.8 (1.90)	26.0	-5.8 (1.83)
Simpson 2000	37.6	-0.5 (1.45)	37.2	-0.5 (1.46)	37.0	-0.5 (1.54)	35.5	-0.5 (1.57)
Pooled Treatment Effect	-2.43 (0.89)		-2.48 (0.89)		-2.42 (0.94)		-2.53 (0.93)	
D-L $\hat{\tau}^2_{\theta_h}$	-		-		-		-	
Counsellor ICC	-		-		0.033		0.033	
Random-Effects Meta-Analysis	**Assuming Independence**				**Allowing for Within-Trial Clustering**			
Level 1 Variance	**Equal**		**Unequal**		**Equal**		**Unequal**	
	(Model 5 with Variance 11)		**(Model 5 with Variance 12)**		**(Model 4)**		**(Model 5 with Variance 15)**	
Trial	**% Weights**	**Mean Difference (Standard Error)**	**% Weights**	**Mean Difference (Standard Error)**	**% Weights**	**Mean Difference (Standard Error)**	**% Weights**	**Mean Difference (Standard Error)**
Chilvers 2001	19.4	0.4 (2.38)	19.3	0.4 (2.40)	19.8	0.4 (2.41)	19.5	0.4 (2.44)
Friedli 1997	26.0	-3.9 (1.75)	25.6	-3.9 (1.79)	25.0	-3.9 (1.92)	25.1	-3.9 (1.91)
King 2000	24.8	-5.8 (1.85)	25.4	-5.8 (1.80)	25.2	-5.8 (1.90)	26.0	-5.8 (1.83)
Simpson 2000	29.8	-0.5 (1.45)	29.7	-0.5 (1.46)	29.9	-0.5 (1.54)	29.4	-0.5 (1.57)
Pooled Treatment Effect	-2.50 (1.40)		-2.52 (1.42)		-2.48 (1.42)		-2.53 (1.43)	
D-L $\hat{\tau}^2_{\theta_h}$	4.50		4.63		4.34		4.48	
Counsellor ICC	-		-		0.033		0.033	

Table 3 Fixed- and Random-Effects Individual-Patient-Data Meta-Analyses of the Absolute Mean Difference in BDI between Counselling and Control where All Trials have Partially Nested Designs

Fixed-Effects Meta-Analysis	**Assuming Independence**		**Allowing for Within-Trial Clustering**	
Level 1 Variance	**Equal**	**Unequal**	**Equal**	**Unequal**
Model	**(Model 16)**	**(Model 17)**		**(Model 18)***
Intercept	16.2 (1.14)	16.3 (1.15)	15.7 (1.19)	15.8 (1.25)
Friedli 1997	-1.3 (1.40)	-1.5 (1.39)	-0.5 (1.58)	-0.5 (1.66)
King 2000	-0.6 (1.38)	-1.0 (1.38)	0.3 (1.49)	0.1 (1.54)
Simpson 2000	0.8 (1.32)	0.7 (1.31)	0.9 (1.42)	0.9 (1.48)
Pooled Treatment Effect	-2.47 (0.90)	-2.47 (0.90)	-2.43 (1.08)	-2.46 (1.12)
$\hat{\sigma}^2_{u1}$			9.66 (6.04)	12.53 (6.41)
$\hat{\sigma}^2_{e0}$	92.77 (6.12)	102.91 (9.47)	89.10 (6.07)	102.20 (9.41)
$\hat{\sigma}^2_{e1}$		82.11 (7.76)		73.20 (7.45)
Counsellor ICC	-	-	0.098	0.146
-2 Log Likelihood	3384	3381	3382	3377
Random-Effects Meta-Analysis	**Assuming Independence**		**Allowing for Within-Trial Clustering**	
Level 1 Variance	**Equal**	**Unequal**	**Equal**	**Unequal**
Model				**(Model 19)***
Intercept	15.5 (1.31)	15.5 (1.37)	15.4 (1.29)	15.4 (1.36)
Friedli 1997	-0.2 (1.74)	-0.3 (1.80)	0.0 (1.73)	-0.1 (1.82)
King 2000	1.0 (1.68)	0.9 (1.74)	1.1 (1.66)	1.0 (1.74)
Simpson 2000	0.8 (1.61)	0.8 (1.68)	0.9 (1.59)	0.9 (1.67)
Pooled Treatment Effect	-2.47 (1.42)	-2.47 (1.42)	-2.49 (1.45)	-2.51 (1.45)
$\hat{\tau}^2$	4.80 (4.58)	4.83 (4.46)	3.85 (4.89)	3.56 (4.76)
$\hat{\sigma}^2_{u1}$			8.06 (6.09)	11.20 (6.54)
$\hat{\sigma}^2_{e0}$	91.88 (6.09)	101.88 (9.38)	88.98 (6.06)	101.87 (9.38)
$\hat{\sigma}^2_{e1}$		81.33 (7.75)		73.24 (7.46)
Counsellor ICC	-	-	0.083	0.133
-2 Log Likelihood	3385	3383	3383	3378

Table 4 Meta-Regression Analyses of the Mean Difference in BDI

Model	Source of Heterogeneity in ICCs	
	Treatment Standardisation (Model 22)	**Patient Eligibility (Model 23)**
Intercept	15.8 (1.27)	15.7 (1.16)
Friedli 1997	-0.6 (1.38)	-0.6 (1.54)
King 2000	-0.2 (1.45)	-0.5 (1.51)
Simpson 2000	0.9 (1.59)	0.9 (1.39)
Counselling	-3.58 (0.91)	-2.37 (1.05)
$\hat{\sigma}^2_{u1}$		8.62 (5.22)
$\hat{\sigma}^2_{e1}$ (Mixed)		80.71 (9.31)
$\hat{\sigma}^2_{e1}$ (Depression)		53.32 (10.92)
$\hat{\sigma}^2_{e0}$ (Mixed)		86.91 (9.32)
$\hat{\sigma}^2_{e0}$ (Depression)		142.06 (25.52)
$\hat{\sigma}^2_{v}$ (Not Standardised)	28.19 (14.07)	
$\hat{\sigma}^2_{v}$ (Covariance)	-15.14 (7.03)	
$\hat{\sigma}^2_{e1}$	71.71 (7.12)	
$\hat{\sigma}^2_{e0}$	102.13 (9.40)	
Counsellor ICC (Mixed)		0.097
Counsellor ICC (Depression)		0.139
Counsellor ICC (Not Standardised)	0.282	
Counsellor ICC (Standardised)	-0.030	
-2 Log Likelihood	3364	3368

Note: Model 22 has been re-parameterised to allow for a negative counsellor ICC for the standardised treatment trials. Model 23 has been adapted to allow for different ICCs for trials with mixed and depression patient eligibility where all trials have partially nested designs.