Analysing Variability in Study Effect among Trials

H. Malcolm Hudson,[1,*] **Victor DeGruttola**[†] **,**[2] **Carol Hargreaves and Val Gebski**[3]

[1]Department of Statistics, Macquarie University, NSW 2109, Australia

[2]Department of Biostatistics, Harvard School of Public Health

[3]NHMRC Clinical Trials Centre, University of Sydney

August 30, 2004

Summary

Meta-analytic approaches that assess sources of variation can enhance the usefulness of non-randomized evidence in a cross-design synthesis. In this paper we estimate heterogeneity in treatment effects associated with different classes of controlled clinical trials. Variance component models are provided for analysis of reported summary statistics when individual case data may not be available. The methodology assists the appropriate weighting of studies, whether randomized or observational, when over-dispersion of trial effects is observed. We evaluate sources of variation among controlled trials in estimates of elevation of risk of invasive breast cancer following hormone replacement therapy (HRT). Variation in results among case-control, prospective, and other non-randomized results is examined and comparison made with effects in two recent large randomized clinical trials.

[*] *email:* Malcolm.Hudson@mq.edu.au

[†] Supported by grant AI51164-01 from NIAID. Visiting NHMRC CTC

Keywords: bias, meta analysis, over-dispersion, random effects, variance components

1. Introduction

While randomized evidence remains the gold standard for comparisons of medical interventions, there is often a considerable amount of non-randomized evidence that could contribute information for such comparisons. The problem with making use of such information is that the choice of treatment in observational studies may be confounded by other factors that impact prognosis, such as disease stage, access to care or demographic factors. The consequence of this confounding is a bias in estimates of treatment differences. While such biases are difficult to identify from any individual study, meta-analytic approaches that combine evidence across a range of studies, both randomized and observational, make it possible to test for the presence of specific effects and therefore enhance the usefulness of non-randomized studies in estimating treatment effects.

There are many sources of variability in estimated study effects. Selection and treatment allocation biases introducing confounding in an observational study will result in the study estimating the wrong treatment effect, while in randomized controlled trials sources of variability in the population recruited to the study and bias due to lack of treatment concealment and differential dropout can introduce specific study effects. Limiting the trials selected for meta review – by varying selection criteria – may limit the introduction of bias but may also lower precision by restricting evidence available. A classic variance-bias tradeoff results.

Schultz, Chalmers, Hayes and Altman (1995) determined a 41% exaggeration of treatment effect in the absence of treatment concealment. Levels of compliance and dropout may differ in populations recruited to studies. Egger and Smith (1997) and Sterne, Gavaghan and Egger (2000) demonstrate other specific study effects introducing variability in the treatment effect: differences in underlying risk of the outcome and size and quality of the study.

This paper investigates meta-analytic approaches for combining evidence across studies when only study-level data (but not individual patient-level data) are available. While access to patient-level data provides the most information for combining across studies, obtaining such data may be difficult and the resources required to create large unified databases from independent studies can be prohibitive.

Our approach is based on the notion that information about variability across studies comes from two different sources. Information about variability within trials is provided by the standard errors of the treatment effect estimates; while information about variability across trials is provided by the variances in the estimated treatment effects across these studies. Identifying sources of variation in reported outcomes and appropriate adjustment for specific variation is conducted within a mixed effects model formulation for observed treatment differences. These models include a random effect of treatment and both fixed and random terms for specific effects associated with the study class.

Section 2 presents the linear mixed effects (LME) models and discusses the conditions required for identifiability. Section 3 summarizes approaches for inference and estimation. Section 3.2 describes the estimation procedures for model parameters and Section 4 provides a data example.

2. Models

2.1 *Stratified analysis of variability in treatment effect by study*

It is useful to begin by considering measurements on each individual in the "raw" data sets available. The implications for use of summary statistics Y substituting for the full analysis will then be discussed.

Specify a mixed-effects meta-analysis model:

$$\begin{aligned} Y_{jk} &= \mu_1 + b_{j1} + e_{jk}, \quad k = 1, \ldots, n_{j1} \\ &= \mu_2 + b_{j2} + e_{jk}, \quad k = n_{j1} + 1, \ldots, n_{j1} + n_{j2} \end{aligned} \tag{1}$$

for $j = 1, \ldots, m$[1]. Here Y_{jk} is the response in individual k in study j, μ_1 and μ_2 are fixed effects, b_{j1}, b_{j2} are random effects associated with the specific trial j (due to variation from other studies in responses measured in its study population and any biases present in that measurement) and with each arm of this study and e_{jk} is the random variability associated with the response of individual k. Of interest is the average treatment effect, $\delta = \mu_2 - \mu_1$. However the observed treatment difference is not only affected by random measurement variability, but also other sources of variation, including potential biases.

The trial specific variations from the average treatment effect δ are random effects, $u_j = b_{j2} - b_{j1}$. These trial specific effects are assumed to be a random sample from a normal distribution with mean 0 and variance σ_1^2.

[1] This model may be expressed as $Y_{ijk} = \mu + \delta_i + b_{ij} + e_{ijk}$, where $i = 1, 2$ are the two arms of the study

Specific effects are independent of the sampling errors $\{e_{jk}\}$. Errors e_{jk} in study j are assumed to have mean 0 and variance σ_{0j}^2.

Observations in study j are Normally distributed with estimate $Y_j = \bar{Y}_{j2} - \bar{Y}_{j1}$ of the treatment effect δ available from each study dataset.

In many meta analyses (e.g. those based on systematic reviews of the literature), 'raw' data Y_{jk} is *not* available but summary measures, such as the means Y_j, can be analysed. Under model (1) above,

$$Y_j = \bar{Y}_{2j} - \bar{Y}_{1j} = \delta + u_j + e_j, \tag{2}$$

where $e_j = \bar{e}_{j2} - \bar{e}_{j1}$, the error in the estimated treatment effect in study j, is distributed $\mathrm{N}(0, v_{0j})$. Here δ is an overall effect of treatment, with random effect u_j varying the treatment comparison due to specific study effects, and $v_{0j} = \sigma_{0j}^2 \left(1/n_{j1} + 1/n_{j2}\right)$ is the variance of the estimate of treatment effect in study j.

In equation (2) δ is the expected treatment benefit. This expected treatment benefit and the between study variability σ_1^2 are of interest; pooled estimation of δ using weighted least squares is natural.

Various assumptions in the model above will need further generalization for more complex circumstances. More complex models involve some or more of the following features: more variance components; study stratification to allow for different treatment effects (still uniform within strata); other study level covariate information defining confounders. We will consider additional variance components in Section 2.5.

The model above may be applied within individual strata (subgroups of trials) within a meta analysis, e.g. within strata comprising randomized and non-randomized studies or within strata comprising randomized studies of low and high quality, however defined. Following stratification, the model terms δ, u_j, e_j are all specific to the strata and it will be of interest to test the homogeneity of δ in different strata.

Models (1) and (2) are special cases of well known general models Harville (1977), Laird and Ware (1982), Laird, Lange and Stram (1987) with linear mixed effects. Stram (1996) provides a framework for the meta-analysis of published data and generalizes some earlier models. Stram's model decomposes the variation in effects measured in different studies into sampling and non-sampling sources. See also Searle et al. (1992, Chapter 6) and McLachlan and Krishnan (1997), Chapter 5.9, for reviews of variance component methods applicable in meta-analysis contexts.

Two representations of random effects models are available as generalizations of the model (2). These forms are the Laird-Ware form

$$y_j = X_j\beta_j + Z_jb_j + e_j \tag{3}$$

and the random effects model form adopted by Searle:

$$y = X\beta + \sum_{i=1}^{p} Z_iu_i + e \tag{4}$$

involving matrices X and Z of known constants. While similar, these forms are not equivalent, Searle's form being more convenient for our purpose. These generalizations permit inclusion of further adjustments for covariates with both fixed and random effects, as will be illustrated in later sections.

With small studies it may be necessary to pool sample variances of small studies. Pooled variance estimates are based on an assumption $v_{0j}^2 = \sigma_0^2\, r_j$, with $r_j > 0$ known. It will usually suffice to make a plausible assumption specifying the relative precision of small studies. For example, we might assume variance homogeneity of small studies. A pooled estimate of variance will then be available with pooled degrees of freedom. When pooled degrees of freedom are large, σ_0^2 can be assumed known.

Within a stratum (such as non-randomized trials, or studies of a common assessed level of 'quality') maximum likelihood (ML) or restricted ML (REML) estimates of $(\mu, \sigma_1^2, \sigma_0^2)$ are available. As noted above, these parameters are permitted to vary between strata, e.g. they may differ between randomized and non-randomized strata.

2.2 *Identifiability*

No matter how large the trials, mean effects alone carry very little information (roughly 1 degree of freedom per trial) concerning variance components. Only their common expected value δ and the variances

$$V_j = \mathrm{Var}(Y_j) = \sigma_{0j}^2(\frac{1}{n_{j1}} + \frac{1}{n_{j2}}) + \sigma_1^2$$

are directly estimable from the means alone. For example, when $Y_1, \ldots, Y_N$ share a common variance, the variance σ_0^2 of the estimated effect and random effect variance σ_1^2 are separately non-identifiable (though their sample variance always provides an upper bound for the between-study variation σ_1^2). The means alone generally provide a poor estimate of σ_1^2.

Additional data, the standard errors and sample sizes commonly available in systematic reviews, resolve any non-identifiability by specifying variances v_{0j} of e_j from within-trial estimates of variance.

Published estimates of treatment differences and their standard errors provide these summary statistics which improve the precision of estimating random effects. If the method of determining confidence intervals is known, standard errors not stated can usually be determined from confidence intervals.

2.3 *Stratified model with odds ratios*

Random effect models can readily be adapted for use with event count response data DerSimonian and Laird (1986). Consider dichotomous outcomes in two trial arms. Then, setting Y_j as the log odds ratio of positive response in the two arms, a confidence interval for the trial specific log odds ratio $\log \omega_j$ is commonly reported. This interval readily provides the point estimate Y_j and its variance v_{0j}.

Model (2) is then applicable, with $k = 1$ and $\text{Var}(e_j) = v_{0j}$ known.

An alternative model proposed in Begg and Pilote (1991) adopts a random baseline common to the two study arms with treatment effect being a fixed effect across studies, unlike the variation in treatment effect specified in the model of der Simonian and Laird.

2.4 *Stratified models: testing homogeneity of study effect*

Model (2) permits stratified analysis of results, as in meta analysis including randomized and non-randomized subgroups of trials. The analysis is repeated separately in subgroups.

Thus the mean and the variance of specific effects may be estimated for trials within each stratum. The lack of fit of this model may be compared with that of a single fit to all studies and a formal test of heterogeneity of means and/or variance components obtained. (The likelihood ratio test for a comparison of models is described in a later section).

When evidence of heterogeneity between subgroups is evident we consider further models intermediate between homogeneous and stratified assumptions, as such models permit pooling of information from the subgroups.

2.5 *Intermediate models: common study effects*

Were study effects to vary randomly around the same mean ($\mu = \delta_1 = \delta_2$) in both randomized and non-randomized subgroups, in which variances of the specific effects differed, then a weighted pooled estimate would better estimate the true effect δ. Restrictions such as this introduced in the stratified model can permit the non-randomized evidence to be informative in estimating the true effect of treatment.

The true effect of treatment may be assumed identical with the parameter μ in a stratum designated the gold standard, perhaps a class of randomized clinical trials, for example. The additional assumption is required to relate population parameters among other strata (such as non-randomized trials) to the corresponding parameters (μ, σ_1^2) in the designated stratum. Without some relationship, other strata cannot be informative.

In general we would expect the variability in the non-randomized studies to be greater than that for randomized studies, because only randomized studies control for confounding factors. If the impact of confounding factors differs across non-randomized studies then this effect would add to the variability in such studies. However, it would also be possible for the variability in non-randomized studies to be smaller than for the randomized studies. If, for example, non-randomized studies were more homogeneous in their study populations (and therefore in factors that independently predict the clinical endpoint, or interact with treatment in affecting the development of the clinical endpoint) than were the randomized studies, this effect could reduce the variability in the non-randomized studies. In addition, if the non-randomized studies were consistent in the effect of the confounding factors, this consistency would introduce a bias relative to the non-randomized studies, rather than an increased variability.

For example, suppose we are considering the relative benefit of two treatments, one of which is more expensive outside of the randomized trial setting. Suppose that in the all of the non-randomized studies people with higher incomes and better access to medical treatment were more likely to receive the more expensive treatment under study. This effect might well induce a bias in the estimated treatment effect, but would not increase the variability across non-randomized studies. In fact this effect could even reduce this variability in the non-randomized studies compared to those that were randomized, if income also is a predictor of the clinical endpoint.

In settings where we do expect (or detect) greater variability in some strata (e.g. non-randomized studies) the expected greater variability in estimating treatment effects in any other stratum we can introduce additional *random* effects specific to strata other than the reference stratum. Each extra random effect introduces higher variability (a variance inflation) in treatment differences in strata in which it is included.

To illustrate one such plausible model assumption relating different strata, consider the specification of equal expected treatment effects ($\mu = \delta_1 = \delta_2$) in randomized and non-randomized populations of trials:

$$\begin{aligned} E(y_j|u) = \mu + u_{j1}, \quad \text{so } y_j \sim \quad & N(\mu, \, \sigma_1^2), \quad \text{for } j \in R \qquad (5) \\ E(y_j|u) = \mu + u_{j1} + u_{j2}, \quad \text{so } y_j \sim \quad & N(\mu, \, \sigma_1^2 + \sigma_2^2), \quad \text{for } j \in NR. \end{aligned}$$

Here R indexes the designated stratum, randomized trials, and NR another stratum, here non-randomized trials. Measurement variances v_{0j}, as before, are assumed known.

Note the assumption introduced that the *expected* bias in non-randomized studies is zero; other random effects models could be considered that utilize other knowledge in forming model assumptions about average specific effects (or biases) in other classes of trials (or postulate relationships between variances of random effects).

Equation (5), which introduces a second random effect component (u_{j2} above), represents a development of model (2). The constrained model is a specific case of the variance component formulation described in Searle et al. (1992). Here

$$y = X\mu + \sum_{i=1}^{2} Z_i u_i + e,$$

where $u_1 = (u_{11}, \ldots, u_{1N})^T$ and $u_2 = (u_{2N_1+1}, \ldots, u_{2N})^T$, X is an arbitrary design matrix for fixed effects, Z_1 is an NxN identity matrix and $Z_2 = \left[0^T \vdots I\right]^T$ is $N \times N_2$, with $N = N_1 + N_2$. Here Z_2 is a partitioned design matrix with a block of zeros of dimension N_1xN_2 for the N_1 studies in stratum 1, and an identity matrix of dimension N_2 corresponding to studies in stratum 2. The error vector $e \sim \mathcal{N}(0, V_0)$ with V_0 diagonal with elements v_{0j}, $j = 1, \ldots, N$.

Upon standardizing the outcomes ($y^\star = V_0^{-1/2} y$), this model is represented in a standard form – Searle et al. (1992), equation (6.2) – with observations with measurement error variance 1.

Another class of models are those of Laird and Ware (1982). The immediate generalization of model (2) for patient level data in this class is:

$$y_j = X_j\mu + Z_j b_j + e_j, \quad \text{for } j = 1, \ldots, M = (N_1 + N_2).$$

For $j = 1, \ldots, N_1,\ N_1 + 1, \ldots, N_1 + N_2$; let y_j be the n_j-vector containing observations on effect in study j. Set $X_j = 1_{n_j}$, a vector of ones, μ to be a scalar, $b_j = (b_{j1}, b_{j2})^T$ and Z_j to be a matrix of dimension $M\,\mathrm{x}\,2$. Also set each of the first N_1 rows of Z to $[1, 0]$, and the remaining N_2 rows to $[1, 1]$. The random effect b_{j2} is an additional variance component. In non-randomized trials this effect inflates variability among this group of trials.

Adopting the standard assumptions of the Laird-Ware model

$$\begin{aligned} \mathrm{var}(e_j) &= \sigma_0^2\, r_j\, I_{n_j} & (6) \\ \mathrm{var}(b_j) &= D, & (7) \end{aligned}$$

where I_n denotes the $n\mathrm{x}n$ identity matrix and D is a 2x2 covariance matrix with variances $\sigma_1^2,\ \sigma_2^2$ and covariance σ_{12}^2. Estimation of (μ, D) will provide the true effect and variances by fixing relative weights applicable to the individual study effects observed in different trials.

Note that the Searle and Laird-Ware formulations differ in that the Laird-Ware model introduces unnecessary and inestimable random effects b_{j2} for studies j in the reference stratum. Searle's random effects model and EM algorithm allow numbers of random effects in the first and second random components to differ.

The Laird-Ware model is often formulated with a general non-diagonal covariance matrix D. However, in this context, direct inspection of the likelihood demonstrates that the structure of D is non-identifiable. Only σ_1^2 and

$\sigma_2^2 + 2\,\sigma_{12}$ is estimable. In these circumstances, it is convenient for us to assume independence of the random effects, as we do hereafter.

3. Inference and Estimation

In any of the models presented above the log-likelihood L is conveniently expressed as

$$-2L = \sum_j \log(V_j) + \sum_j \frac{(y_j - \mu)^2}{V_j}$$

where V_j represents the variance of the treatment summary outcome in trial j according to the model. For example, in Searle's formulation:

$$\begin{aligned} V_j &= v_{j0} + \sigma_1^2 \qquad \text{for } j = 1, \ldots, N_1 \\ &= v_{j0} + \sigma_1^2 + \sigma_2^2 \qquad \text{for } j = (N_1 + 1), \ldots, N \end{aligned} \tag{8}$$

3.1 *Inference*

Nested models may readily be compared by comparison of log-likelihoods, once variance parameters are estimated. The resulting difference in twice log-likelihood, denoted by $-2\Delta l$ in our applications, should be compared with *half* the tabled value for chi-square with degrees of freedom the number of extra variance parameters; see Stram and Lee (1994).

3.2 *Estimation*

3.2.1 Laird-Ware model form The Maximum Likelihood (ML) for model (2) estimators may readily be obtained using the EM algorithm of Laird and Ware (1982) for the mixed effect model

$$y_j = X_j\beta + Z_j b_j + e_j, \quad \text{for } j = 1, \ldots, N, \tag{9}$$

where b_j, corresponding to u_j in (2), is a random effect for observations indexed by j. Here parameters β and $\{b_j\}$ refer to a single stratum; estimation may be repeated within each stratum.

Equation (2) is readily reduced to a canonical form with observations of equal variance. Setting $y_j = Y_j/\sqrt{r_j}, \beta = \delta, X_j = 1/\sqrt{r_j}, Z_j = 1/\sqrt{r_j}$ and substituting $\tilde{e}_j = e_j/\sqrt{r_j}$ the model becomes (9) with $u_j \sim N(0, D)$ and $\tilde{e_j} \sim N(0, \sigma_0^2)$.

A similar rescaling is required for the case of binary outcomes, where der Simonian's approach adopts the same model for the log odds-ratio in study j.

The application of an EM algorithm to this canonical Laird-Ware form is standard.

With more general models involving additional random effects, EM estimation is again available once the design structure is specified by choice of design matrices X_j and Z_j. The interested reader is directed to McLachlan Ch. 5.9.1.

3.2.2 Searle's random effects model Searle's model form is readily adapted to models with either one or two random effect components in each study. The use of an EM algorithm in the fitting of parameters in this class of model is described in Searle, Chapter 8.3. The application of the EM algorithm from first principles is straightforward in this context. EM specifies weighted least squares estimation of μ and weighted residual sums of squares estimates for σ_1^2 and σ_2^2 using weights inverse to the current estimates of the variances defined in equation (8).

3.2.3 Prediction Best linear unbiased prediction (BLUP) of individual random (specific) effects is readily available in variance component models. Thus improved estimates of odds ratios for individual studies may be provided. Our approach is described in Searle, Chapter 7.4.

4. Relationship between use of hormone replacement therapy and risk of breast cancer

The increased risk of breast cancer resulting from use of hormone replacement therapy (HRT) has been settled only quite recently, and review of the large numbers of studies and the consistency of conclusions reported in the numerous studies are considered of interest.

A systematic review HBFC Group (1997) conducted by the Collaborative Group on Hormonal Factors in Breast Cancer (HFBC) appeared in the Lancet in 1997. This paper presented a meta analysis of results of 51 epidemiological studies of the relationship use of HRT in post-menopausal women and between breast cancer. Since that time two large randomized clinical trials have closed and reported findings while other non-randomized controlled trials have appeared in the literature.

Findings of the review and later studies indicated increased relative risk of diagnosis of breast cancer among current or recent users of HRT.

The meta analysis indicated that this excess risk increases with duration of use but reduces after cessation of use and appears to have largely disappeared after about 5 years. Strong potential for confounding exists in observational studies between the timing of the menopause and use of HRT. Failure to take time since menopause into account leads to substantial underestimation of the risk of breast cancer associated with the use of HRT.

BMI is also a potential confounder. Biases may be present in some studies due to earlier diagnosis and differential reporting of use of HRT. Most (87%) current or recent users in the meta analysis had begun use within 5 years of the menopause and 97% were aged under 70 at the time of breast cancer diagnosis (on average in 1985). Little consolidated information was available about other adverse effects of HRT use.

In 1998 study results Hulley, Grady and other authors (1998) from a randomized trial of Estrogen and Progestin hormone use for secondary prevention of coronary heart disease in post-menopausal women were published. This Hormonal Estrogen receptors study (HERS) followed a total of 2763 women with coronary disease for an average follow-up of 4.1 years, subsequently extended (HERS-II) after unblinding. While hormonal treatment achieved significant reductions in cholesterol over placebo, no overall cardiovascular benefit or significant differences in several other endpoints for which power was limited was observed. The latter endpoints included breast cancer, with 32 and 25 events, providing estimate of relative risk 1.30 with confidence interval (CI, 0.77- 2.19).

Women's Health Initiative (WHI) Investigators reported WHI Investigators (2002) on a recent randomized clinical trial of 16,608 post-menopausal women aged 50-79 years with an intact uterus at baseline who were recruited by 40 US clinical centres in 1993-1998. This group was part of a larger study for which the primary outcome was coronary heart disease (CHD). Invasive breast cancer was designated as a primary adverse outcome the reported substudy, which closed early with evidence of adverse effects including breast cancer harm, some increase in CHD, stroke and PE outweighing benefits over the average 5.2 year follow-up period in the estrogen plus progestin arm. The hazard ratio for invasive breast cancer was then 1.26 (95% nominal CI 1.00-1.59, adjusted for multiple outcomes 0.83-1.92) between HRT assignment (n=8506, 231 deaths, 3 with breast cancer cause) and placebo (n=8102, 218 deaths, 2 with breast cancer cause). Adjustments for medication compliance were undertaken, censoring non-compliers 6 months after non adherence, increasing the hazard ratio for breast cancer to 1.49.

Studies can be classified by elements of their design in order to examine specific effects associated with each class. We classified studies as either randomized clinical trials, prospective studies, retrospective studies with hospital controls or retrospective studies with community controls. This classification was consistent with the approach adopted in the systematic review.

4.1 *Data analysis*

Trial data on odds-ratios and standard errors was extracted from the HFBC review appearing in the Lancet, the abovementioned two randomized trials and other later publications Persson, Thurfjell, Bergstrom and Holmberg (1997), Jernstrom, Frenander, Ferno and Olsson (1999), Olsson, Bladstrom, Ingvar and Moller (2001) identified in literature searches. The systematic review was conducted prior to the reporting of two recently conducted large randomized clinical trials. These and results of two further studies (Swedish cohort and case-control studies) on the scientific question were included in the analysis reported below.

The preferred outcome and exposure comparison in the odds-ratios reported compares users or recent HRT users with non-users (in the 5 years previous). Recall that the mean time of follow up following randomization in the Women's Health initiative was just over 5 years and that during that period of time 30% of the study participants failed to comply with their treatment regime. Collection of recent usage statistics within each study involves a complex time dependent analysis, for which source data was not available. Instead, we chose for analysis the odds ratio comparison of ever-use with never-use. Suitable data for this analysis was obtained at study level from the meta analysis comparison of studies conducted by the HBFC (see their Figure 3) and in subsequent study publications.

The odds ratio (or hazard ratio) in each study was supplemented by its reported standard error. Standard errors were calculated for log odds-ratios from confidence intervals or other statistics reported in source papers. In the case of the WHI trial, it was considered appropriate to base calculations on the standard error adjusted for multiple outcomes and sequential stopping criteria, as reported in that study.

We report below results of random effects models conducted with routines developed in S-Plus 6.1.

We applied the linear mixed effects model of Searle et al. (1992) according to the EM algorithm of Chapter 8 with straightforward modifications allowing for (i) known, but (ii) unequal, variances of observations.

4.2 *HRT studies results*

There were 28 reported studies (see Table 5) providing hazard or odds ratios for ever- versus never- use of HRT.

[Table 1 about here.]

When meta analysis was applied to these 28 studies the average log ratio (log-OR) was 0.186 with weighted standard deviation 0.165. Weighting was by study sample size (appropriate in a population of homogeneous trials). Figure 1 displays the individual odds ratios and their 95% confidence intervals as reported. The Figure suggests heterogeneity of study outcome is present, as not all confidence intervals appear consistent with the pooled OR estimate, 1.20.

A linear mixed effect (LME) model provides for variation in odds ratios through introduced random effects. Table 5 shows parameter estimates, weighted residual sums of squares (-2 log-Likelihood) and corresponding degrees of freedom for three models: the null model just described, an LME with common heterogeneity in non-randomized studies of all classes and an LME with heterogeneity in both RCTs and non-randomized clinical trials.

[Table 2 about here.]

It appears from these results that while there is significant evidence of heterogeneity among non-randomized trials there is no evidence of lack of consistency in findings of the two recent RCTs. In either model the pooled estimate of the log-OR $\hat{\mu}$ is 0.188, again suggesting a 21% excess risk of invasive breast cancer on average.

It is of interest to identify whether particular classes of non-randomized studies exhibit specific effects that distinguish them from other classes.

[Figure 1 about here.]

When the trials are grouped according to class of study, as in Figure 1, there appears to be a reasonable consistency overall with the log-OR estimate 0.188, with a few exceptions. Firstly, in the group of prospective studies, the result from pooling other (small) prospective studies differs substantially from the others. The pooled log-OR for 'other prospective studies' is -0.48 (RR 0.62). Population based case-control studies provide one study discordant with the general finding, the Stanford study with log-OR estimate -0.29 (RR 0.75), again suggesting reduced risk of invasive breast cancer. This group also has a number of confidence intervals not including the common estimate. While there is no evidence of overall bias in this group of studies, the results are consistent with a small random effect creating extra variation of results in this class, leading to underestimation in confidence intervals of the range of values for a common population mean. That is, standard errors may be underestimated in this group of studies. Finally, among the few hospital based case-control studies, one study (the LaVecchia study) substantially *over*-estimates the common log-OR.

It is possible to identify many of the trials providing discordant results by the discrepancy observed in these trials between the variance of the log odds-ratio for a pooled cross-classification (ignoring stratification) and the variance reported after adjustment for stratification. This is a measure of imbalance in the study design.

The general finding appears to be that all groups provide an overall average that is consistent between study types, but that non-randomized studies, and case-control studies in particular, exhibit occasional studies with specific effects consistent with a mixed effects model. The random effects may have greater variances in these classes.

[Figure 2 about here.]

According to the model developed, a 95% confidence interval for the log odds-ratio comparing ever use with never use of HRT in post-menopausal women is the range 0.16 to 0.22.

5. Conclusion

This paper contributes a random effects mixed model methodology for calculating appropriate weightings of evidence contributed by different classes of study in meta analysis. Droltcour et al. (1993) has argued for the necessity of cross-design synthesis for combining RCT and medical-practice databases. Their recommended approach comprises: identifying complementary research designs and studies conducted; in depth assessment of each study to identify the chief potential biases associated with its design; 'secondary adjustments' of study results to correct known biases; developing synthesis frameworks and models to minimize the impact of hidden biases. We have offered a general modeling methodology as the necessary synthesis framework. Secondary adjustments may also be made by inclusion of the covariate information necessary to such bias adjustment.

In the application to HRT data we observed significantly greater variability in risk associated with use of Estrogen/Progestin in different studies than is consistent with a single meta effect, despite the consistency in findings of two recent randomized clinical trials. Applying the linear mixed effect model which introduces additional variation in studies other than these two RCTs led to two benefits: (i) providing a better fitting model for all study results with corresponding pooled meta effect, and (ii) a reduction in the uncertainty of the risk estimate from the findings of the RCTs alone. No single class of non-randomized study exhibited either consistent bias generally increasing or reducing risk estimates based on that study class or higher random effects variability than other classes.

This is not to say that data would not be consistent with other assumptions, such as a population of studies which are generally homogeneous, with the exception of a few ‘outlier’ studies that should carry very little weight. A slight variant on the variance component model we have applied, e.g. fitting a non-normal distribution with extra weight in its ‘tails’, will provide appropriate meta parameter estimates in this case. In our application to HRT data, however, inspection of Figure 1 provides little sharp evidence against the normality assumption.

It is also possible to simply generalise our models to adjust for other fixed effects or allow for situations where extra variation is not apparent in non-randomized populations.

It should be noted that the odds ratios from individual non-randomized trials employed in our HRT modeling were those adjusted for a number of important potential confounders in the Lancet meta-analysis. This is a very good practice that would be expected to substantially reduce potential biases within an observational class of studies. This adjustment was particularly assisted by the comprehensive review of potential confounders in that paper, meaning that all odds-ratios were adjusted for the same stratifying factors.

The general consistency of effects observed in development of invasive breast cancer in randomized and non-randomized controlled clinical trials of HRT is not always present. In studies of HRT with regard to purported beneficial effects on coronary heart disease (CHD) outcomes, prevailing observational evidence at the time the WHI and HERS clinical trials were designed was for significant reduction in CHD with use of hormones. The eventual findings of the same randomized clinical trials were of no significant benefit of HRT. While there are no existing meta analyses of the evidence collected in non-randomized studies, in this case it is likely that the evidence of the two classes of study were in conflict, so another source of variation should exist. This raises the interesting question of why analysis of the primary designated positive outcome (CHD) is confounded but analysis of the primary adverse outcome, invasive breast cancer, is not.

Acknowledgements

The authors would like to thank Sally Lord for her assistance in the literature review.

References

Begg, C. B. and Pilote, L. (1991). A model for incorporating historical controls into a meta-analysis. *Biometrics* **47**, 809–906.

DerSimonian, R. and Laird, N. (1986). Meta analysis in clinical trials. *Controlled Clinical Trials* **7**, 177–188.

Droltcour, J., Silberman, G. and Chelimsky, E. (1993). Cross-design synthesis. *International Journal of Technology Assessment in Health Care* **9**, 440–449.

Egger, M. and Smith, G. (1997). Meta-analysis. potentials and promise. *British Medical Journal* **315(7119)**, 1371–4.

Harville, D. A. (1977). Maximum-likelihood approaches to variance component estimation and to related problems. *J. Amer. Stat. Assoc.* **72**, 320–340.

HBFC Group (1997). Breast cancer and hormone replacement therapy: collaborative reanalysis of data from 51 epidemiological studies of 52,705 women with breast cancer and 108,411 women without breast cancer. *Lancet* **350**, 1047–59. [Prof Valerie Beral, ICRF Cancer Epidemiology Unit].

Hulley, S., Grady, D. and other authors (1998). Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. *Journal of the American Medical Association* **280**, 605–613.

Jernstrom, H., Frenander, J., Ferno, M. and Olsson, H. (1999). Hormone replacement therapy before breast cancer diagnosis significantly reduces the overall death rate compared with never use among 984 breast cancer patients. *British Journal of Cancer* **80**, 1453–8.

Laird, N., Lange, N. and Stram, D. (1987). Maximum likelihood computations with repeated measures: Application of the EM Algorithm. *Journal of the American Statistical Association* **82**, 97–105.

Laird, N. M. and Ware, J. H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics* **38**, 963–974.

McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions.* Wiley.

Olsson, H., Bladstrom, A., Ingvar, C. and Moller, T. (2001). A population based cohort study of hrt use and breast cancer in southern sweden.

British Journal of Cancer **85**, 674–7.

Persson, I., Thurfjell, E., Bergstrom, R. and Holmberg, L. (1997). Hormone replacement therapy and the risk of breast cancer. nested case-control study in a cohort of swedish women attending mammography screening. *International Journal of Cancer* **72(5)**, 758–61.

Schultz, K., Chalmers, I., Hayes, R. and Altman, D. (1995). Empirical evidence of bias. dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Journal of the American Medical Association* **273**, 408–412.

Searle, S. R., Casella, G. and McCulloch, C. E. (1992). *Variance Components.* Wiley, New York.

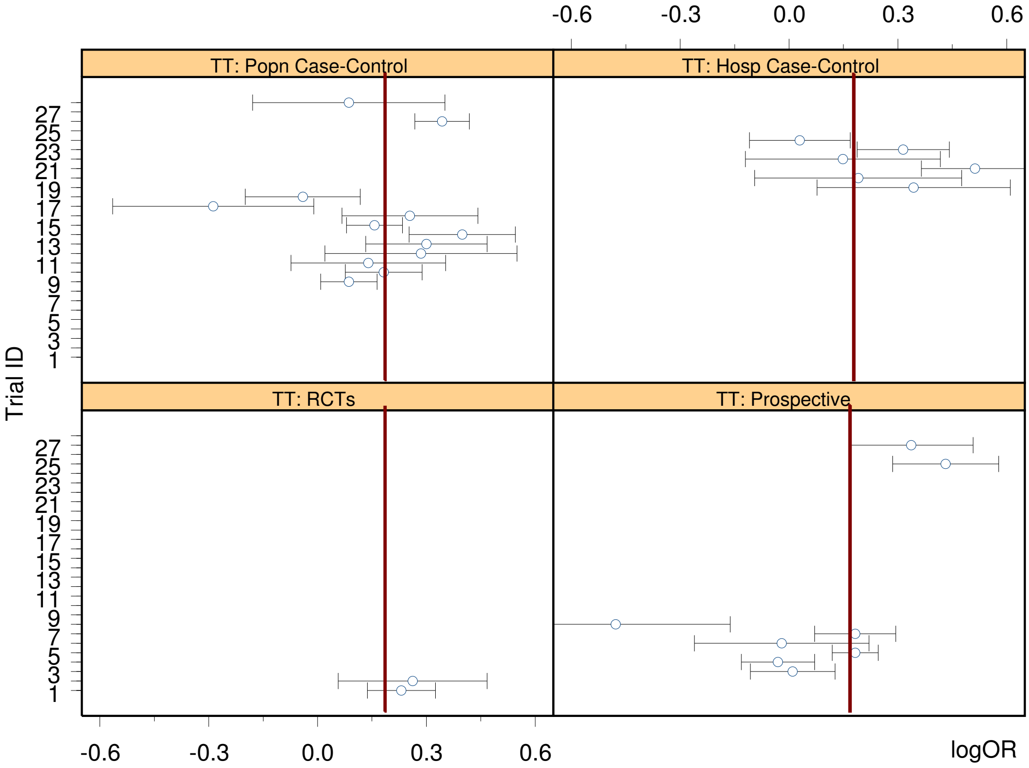
Sterne, J., Gavaghan, D. and Egger, M. (2000). Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology* **53(11)**, 1119–29.

Stram, D. and Lee, J. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics* **50**, 1171–1177.

Stram, D. O. (1996). Meta-Analysis of Published Data Using a Linear Mixed-Effects Model. *Biometrics* **52**, 536–544.

WHI Investigators (2002). Risks and benefits of estrogen plus progestin in healthy postmenopausal women. *Journal of the American Medical Association* **288**, 321–333.

Figure 1. Relative risk or odds ratio and 95% confidence intervals, n=28. Studies ordered by date of publication.



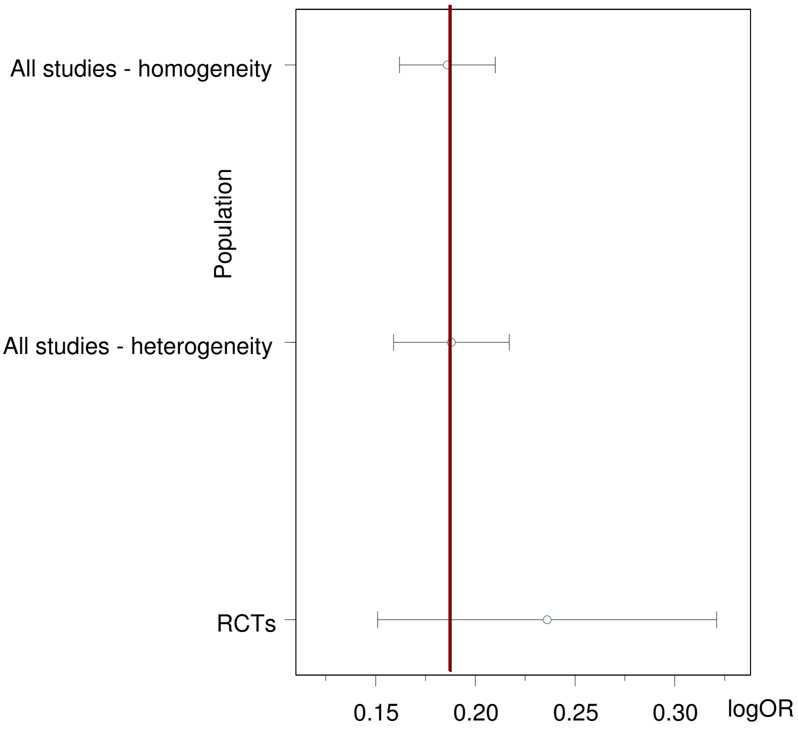


Figure 2. 95% confidence intervals for μ based on different study population models.

Table 1

HRT Trials with year of publication

Year	Trial
2002	Women's Health Initiative(WHI)
1998	Heart and Estrogen/progestin replacement Study (HERS)
1985	Canadian NBSS
1985	Schairer
1986	Nurses' Health
1988	Netherlands Cohort
1991	Iowa Women's Health
	Other Prospective
1976	Brinton
1981	Cash
1981	Hislop
1983	Bain
1983	Ewertz
1984	Long Island
1988	Four State Study
1989	Yang/Gallagher
1989	Stanford
	Other Case Control Pop Controls
1974	Morabia
1982	Vessey
1987	La Vecchia
1990	Katsouyanni
1992	Franeschi
	Other Case Control Hosp Controls
1992	Nurses Cohort
1995	Nurses Cohort Extension
1999	Swedish Cohort
1999	Swedish Case-Control

Table 2

Model estimates and log-Likelihood statistics; after 1000 EM iterations

Model	Parameter estimate	$-2l$	df
Homogeneous model	$\hat{\mu} = 0.186$		
no random effects	$\sigma_1^2 = 0$		
	$\sigma_2^2 = 0$	37.73	27
Heterogeneity			
in non-randomized studies only,	$\hat{\mu} = 0.188$		
shared mean	$\sigma_1^2 = 0$		
	$\hat{\sigma}_2^2 = 0.00684$	27.407	26
Heterogeneity but shared mean	$\hat{\mu} = 0.188$		
in both RCTs and NRCTs,	$\hat{\sigma}_1^2 = 0.00011$		
shared mean	$\hat{\sigma}_2^2 = 0.00672$	27.405	25