Rice, K., Higgins, J. P. T., & Lumley, T. (2018). A re-evaluation of fixed effect(s) meta-analysis. *Journal of the Royal Statistical Society: Series A*, *181*(1), 205-227. https://doi.org/10.1111/rssa.12275

A re-evaluation of fixed-effect(s) meta-analysis

Kenneth Rice *[1], Julian PT Higgins[2,3] and Thomas Lumley[4]

[1]Department of Biostatistics, University of Washington
[2]School of Social and Community Medicine, University of Bristol
[3]Centre for Reviews and Dissemination, University of York
[4]Department of Statistics, University of Auckland

January 9, 2017

Abstract

Meta-analysis is a common tool for synthesizing results of multiple studies. Among methods for performing meta-analysis, the approach known as 'fixed-effects' or 'inverse-variance weighting' is popular and widely-used. A common interpretation of this method is that it assumes the underlying effects in contributing studies are identical, and for this reason it is sometimes dismissed by practitioners. However, other interpretations of fixed-effects analyses do not make this assumption, yet appear to be little known in the literature. In this paper, we review these alternative interpretations, describing both their strengths and limitations. We also describe how heterogeneity of the underlying effects can be addressed, with the same minimal assumptions, through either testing or meta-regression. Recommendations for the practice of meta-analysis are given; it is hoped that these will foster more direct connection of the questions meta-analysts wish to answer with the statistical methods they choose.

1 Introduction

Meta-analysis is used to combine results from multiple, independent, studies that have similar aims. It is now a widely-established statistical tool (Borenstein et al., 2009); common applications include those to randomized trials of the

*kenrice@u.washington.edu

same comparison of interventions (Higgins and Green, 2011), to estimates of correlation between two variables (Hunter and Schmidt, 2004), and to measures of the accuracy of a diagnostic test (Leeflang et al., 2008). More sophisticated meta-analysis methods are used to understand inter-study differences (Berkey et al., 1995), to compare multiple treatments simultaneously (Lu and Ades, 2004) and, more generally, to synthesize evidence from different types of studies (Ades et al., 2008).

A common procedure in a basic meta-analysis is to calculate a weighted average of parameter estimates from different studies, where the weights are given by the inverses of the estimated variances of these estimates. This approach is typically referred to as a 'fixed-effects' (or 'fixed-effect') meta-analysis, and its use is standard in many fields of application. However, the approach has been criticized, because it does not directly address between-study differences, in particular the possibility that studies' parameter estimates differ by more than chance alone. Because of this, an alternative 'random-effects' meta-analysis is often advocated, in which the weights in the weighted average additionally incorporate a second component of variation that describes the differences in the studies' underlying parameters (DerSimonian and Laird, 1986).

There has been considerable debate over the relative merits of fixed- and random-effects approaches to meta-analysis, without a common consensus being reached about which should be used and when; an example of where this lack of consensus causes real difficulty is given in Section 1.1. In the rest of the paper we aim to inform this discussion, by providing a detailed interpretation of the fixed-effect(s) meta-analysis approach. Section 2 describes how apparently conflicting statistical advice can be reconciled by drawing a distinction between a fixed-*effect* (singular) model and a fixed-*effects* (plural) model. In Section 3 we discuss in detail what the inverse variance weighted average represents, and how it should be interpreted under these different models. In Section 4 we show how methods for assessing heterogeneity can be viewed as natural complements to the fixed-effects analysis, even when they are not being used to test homogeneity. We conclude with discussion of strengths and limitations in Section 5.

1.1 Illustrative Example

An example illustrating the major issues is given in Figure 1, which shows a meta-analysis undertaken in a recent genetic study (Levy et al., 2009). In it, data from the participants in six cohort studies is used to address the linear association between an outcome of interest (the participants' blood pressure levels) and a covariate (the number of copies they have of the 'A' allele at the rs11014166 variant), which takes values between 0 and 2. Testing the null hypothesis of no effect is one important goal here, but estimation of an overall effect size is also

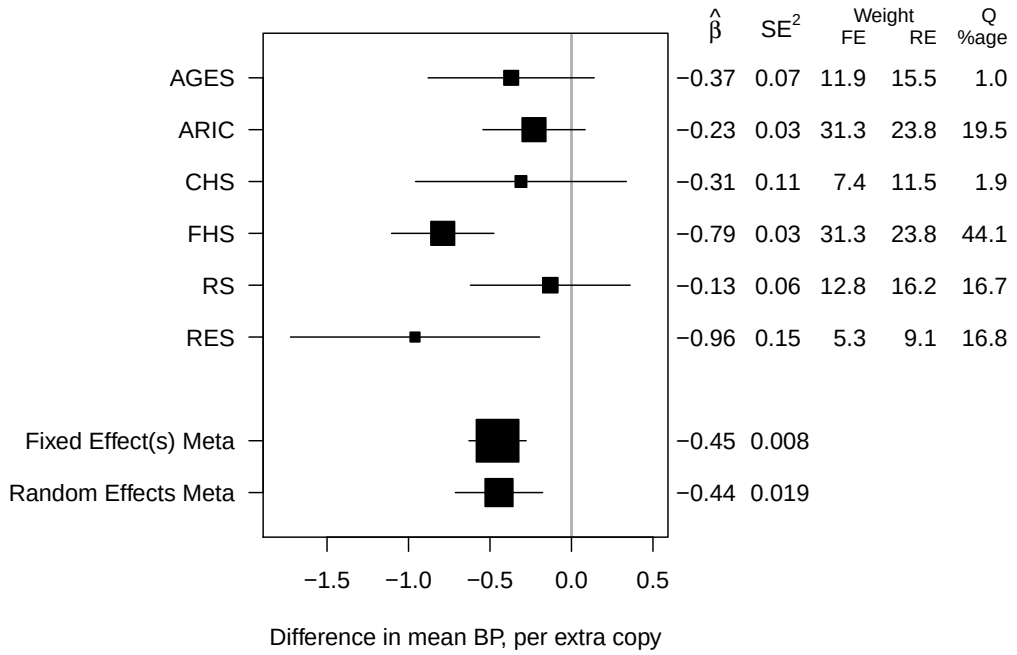


Figure 1: Data and results from a meta-analysis of the association between diastolic blood pressure levels and number of copies of the 'A' allele at variant rs11014166; for details of contributing studies see Levy et al. (2009). Box sizes are proportional to the precision of corresponding study-specific estimates. The weights column indicates those for standard fixed- and random-effect meta-analyses, as formally defined in Section 2. The final column is the relative size of study-specific contributions to Cochran's Q, as a percentage; see Section 5 for definitions.

needed, to assess the magnitude of association and to decide upon next research steps (Zhong and Prentice, 2010). As the figure shows, the decision to use fixed- or random-effects meta-analysis here makes a non-trivial difference to the summary inference; the overall precision in the estimates differs by more than a factor of two, i.e. the difference we might expect to see by changing the sample size by a factor of two. For another example with similar consequences, see Stephens and Balding (2009, pg 689).

A notable feature of Figure 1 is that the effect of the genetic variant shows heterogeneity across the studies. For this data, Cochran's Q statistic (see Section 5) is 10.02, which in the usual statistical test of homogeneity gives a p-value of 0.074, below the $\alpha = 0.1$ threshold for significance recommended by some au-

thors (Berman and Parker, 2002). Similarly, the Der-Simonian Laird estimate of the standard deviation of the study effects is 0.23, and the I^2 statistic is 50.1%, which "may represent considerable heterogeneity" (Higgins and Green, 2011, Section 9.5.2). Nevertheless, in the original publication of Figure 1's data, the fixed-effects meta-analysis *was* used, without regard to observed or actual heterogeneity — and for meta-analysis of genomewide association studies, this practice is standard. Using fixed-effects meta-analysis in this way has been noted to be practically useful in this field (Pereira et al., 2009) when, as is typical, we seek novel associations between the outcome and genetic variants. Furthermore, de Bakker et al. (2008) note that random-effects analysis may be "too conservative", and that in this field fixed-effects meta-analysis may be preferred.

Based on the evidence of heterogeneity across studies, many statisticians would rule out using fixed-effects meta-analysis in Figure 1. For example Hartung et al. (2011, pg 63) state that "It is important to make sure that the underlying effect sizes arising out of experiments are indeed homogeneous before performing any meta-analysis". Other texts provide the same advice, rejecting fixed-effects meta-analysis when heterogeneity is observed; for an example see (Lipsey and Wilson, 2001, pg 133). Other authors (Schmidt and Hunter, 2004, pp 199, 397) go further still, arguing for *never* using fixed-effects meta-analysis in practice, regardless of the observed data, on the grounds that true homogeneity is never plausible.

In contrast, other parts of the statistical methods literature portray the fixed-effects meta-analysis as appropriate, when inference is sought on the average association across the studies at hand (Fleiss, 1993, pg 124). The precise meaning of 'average' here is described by Konstantopoulos and Hedges (2009), and a rationale for using fixed-effects meta-analysis even in the presence of heterogeneity is described by Hedges and Vevea (1998). This work can be seen as a formalization of ideas supporting fixed-effects meta-analysis set out by R. Peto (Peto, 1987), who has also advocated for use of this analysis in practice (Early Breast Cancer Trialists' Collaborative Group, 1998).

Situations like those in Figure 1 therefore place applied statisticians in a difficult position. If you, as the statistician for such a study, recommend a random-effects meta-analysis for this data, you must explain to co-authors and peer reviewers why the standard approach in the genetics field is being passed over. Alternatively, if you suggest a fixed-effects meta-analysis, other co-authors and reviewers have ample reason for concern that your analysis is founded on unrealistic assumptions of homogeneity. As the results of the two analyses are importantly different, and only one can be the primary results, one group of co-authors/reviewers must end up dissatisfied.

This incongruous and unattractive situation is not restricted to genetics; the conflicting statistical advice above would apply meta-analyzing data of any form.

Moreover, the lack of a coherent statistical framework, that might be used to decide between fixed- and random-effects meta-analysis in a more satisfactory way, makes it difficult for students to learn the relevant statistical methods — and to understand what the methods can and cannot provide.

2 The inverse-variance weighted average

2.1 Calculations

In this section we review procedures for meta-analysis using inverse-variance weights. Suppose we have k distinct studies, and each supplies its independent estimate $\hat{\beta}_i$ of study-specific parameter β_i, where $1 \leq i \leq k$. Study parameter β_i reflects the true state of nature in population i, from which the data in study i is sampled. Reflecting this sampling uncertainty, each study also supplies $\hat{\sigma}_i$, an estimate of the standard error of $\hat{\beta}_i$. A meta-analysis is to be performed in order to summarize these studies' data, and the summary produced should be interpretable, relevant to the scientific question at hand, and statistically well-calibrated. Typically, the summary comprises a point estimate and confidence interval for some overall measure of effect. (We consider testing null hypotheses of no effect in Section 4.1.)

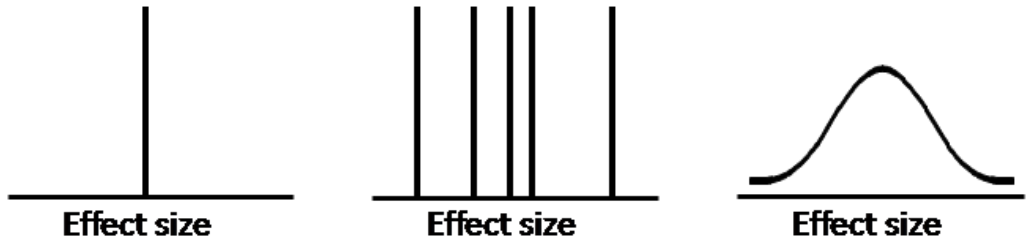
The inverse-variance weighted average is given by

$$\hat{\beta} = \frac{\sum_{i=1}^{k} w_i \hat{\beta}_i}{\sum_{i'=1}^{k} w_{i'}}, \tag{1}$$

where $w_i = 1/\hat{\sigma}_i^2$ (Whitehead and Whitehead, 1991). If each study is large and has implemented a valid analysis, then approximate Normality of the estimates $\hat{\beta}_i$ will hold and each $\hat{\sigma}_i$ will estimate the standard deviation of $\hat{\beta}_i$ with error that is negligible compared to the true standard deviation of $\hat{\beta}_i$. For brevity, we shall refer to these conditions as 'standard' throughout, and we note that they are often assumed implicitly. In meta-analysis, the standard conditions ensure that, to a good approximation, $\hat{\beta}$ is also Normal, and that its standard error may be estimated by

$$SE\left(\hat{\beta}\right) = \sqrt{\frac{1}{\sum_{i=1}^{k} w_i}} \tag{2}$$

with negligible error, compared to the true standard deviation of $\hat{\beta}$ (Whitehead and Whitehead, 1991). That these estimates have the statistical properties we have described is well-known and is not debated, regardless of whether the β_i are all identical, all different, or anything in between. The controversial issue we aim to clarify is whether $\hat{\beta}$ in (1) estimates a parameter of scientific relevance.



(a) Common effect model (b) Fixed effects model (c) Random-effects model

Figure 2: Illustration of three different assumptions possibly relevant to meta-analysis

2.2 Possible assumptions about similarity of effects in different studies

To determine whether the estimate $\hat{\beta}$ in (1) is relevant, we need to consider the modeling assumptions that underlie the meta-analysis. We follow Spiegelhalter et al. (2004, Chapter 3) and distinguish between three assumptions that can be made about the true effects, β_i, across studies. For now, we restrict ourselves to the situation of a collection of studies with similar aims and designs, free of important flaws in their implementation or analysis. (See Section 6 for further discussion.) We illustrate the three assumptions relevant to such a situation in Figure 2.

First, the *identical parameters* (Figure 2a) assumption asserts that all of the β_i are identical, say to β. In other words, exactly the same true effect underlies each and every study. In the genetic example of Figure 1, the statement would be that the association between blood pressure and genotype is equal across all study populations. Some authors refer to this assumption as the *fixed-effect model*, in the singular, although in this paper we use *common-effect model* for clarity. This model is often presented in the literature as the basis for estimation using the weighted average in (1), for example by Hartung et al. (2011, Chapter 4).

Second, the *independent parameters assumption* (Figure 2b) asserts that the β_i are unrelated. In the genetic example of Figure 1, we see how this is a minimal assumption; it would state only that there is an effect of genotype in each study population, with no regard as to whether these effects are similar, or not. We will refer to this assumption as the *fixed-effects model*, in the plural. This model underlies the fixed-effects meta-analyses presented by, for example, Hedges and Vevea (1998).

Third, the *exchangeable parameters* (Figure 2c) assumption treats the β_i as independent identically-distributed draws from a single 'mixing' distribution. In

the genetic example, this means that the sizes of the genotype's effect in different populations, while not necessarily identical, will be somewhat similar; the degree of similarity is governed by the parameters of the mixing distribution, which are fit by the available data. This approach can also be motivated without the i.i.d. assumption; use of a mixing distribution is equivalent to a statement of *a priori* exchangeability (Bernardo and Smith, 2003, Chapter 4.2); in the genetic example, this means only that knowledge of the genotype's effects in different population can not be distinguished *a priori*. Interpreted this way, assuming exchangeable parameters is in between the assumption that all β_i are common, versus all unrelated. However it is motivated, assuming that the β_i represent samples from a mixing distribution is known as the *random-effects model* in meta-analysis. The random-effects model is discussed at length elsewhere (Higgins et al., 2009); we particularly note here that its assumptions are, in general, different from both the common-effect and fixed-effects model.

We have stated these assumptions at length because, in discussing fixed- versus random effects, confusion often arises about the motivation for the weighted average in (1), and its estimated standard error given in (2). As we shall see in the next section, these estimates can be reasonably motivated under *both* the common-effect and fixed-effects models.

3 Interpretation of $\hat{\beta}$ under common-effect and fixed-effects models

As stated in Section 2.1, in meta-analysis of large studies that all report valid analyses, it is not debated that $\hat{\beta}$ in (1) and $SE(\hat{\beta})$ in (2) respectively define an estimate and its variability, under the standard conditions. But exactly what parameter is $\hat{\beta}$ estimating?

Under the common-effect model, the answer is simple: all studies use data from their respective populations to estimate the same common effect, which we denote as β; we note that β reflects the effect in any combination of the individual study populations. Estimator $\hat{\beta}$ in (1) can be interpreted as a weighted average of several estimates of β, and it has several good statistical properties. It is consistent for the common effect, i.e. with large samples the weighted average $\hat{\beta}$ will approach the true value β, and if each $\hat{\beta}_i$ is unbiased for the common effect, it also follows that $\hat{\beta}$ is unbiased for β. Furthermore, from the Crámer-Rao lower bound (Schervish, 1995, Chapter 5), it follows that the inverse-variance weighted estimate will be efficient (i.e. has minimum variance) amongst the class of estimates that are unbiased for β.

Under the fixed-effects model, $\hat{\beta}$ also estimates a population parameter, for the population formed by amalgamating the study populations at hand. The

parameter can be described in terms of three components. First, let β_i denote the effect in study population i, as before. Second, let ϕ_i denote the rate at which (under the standard conditions) statistical information accrues about β_i, per observation from study population i; this quantity is also known as the Fisher information (Casella and Berger, 2001, pg 115) and we note it would be equivalent to define $\phi_i = 1/(n_i\sigma_i^2)$, where σ_i denotes the standard error of $\hat{\beta}_i$ and n_i is the sample size in study i. (We note ϕ_i is, in general, an average rate of information accrual, averaging over the observations from study population i.) Finally, in the overall population that amalgamates all k individual study populations, define η_i as the proportion of the population that comes from study population i.

Using this notation, we can define β as

$$\beta = \frac{\sum_{i=1}^{k} \eta_i \phi_i \beta_i}{\sum_{i'=1}^{k} \eta_{i'} \phi_{i'}}. \tag{3}$$

Parameter β and the population it describes are discussed further in Section 3.1. For now, we see that under the fixed-effects model β is a parameter describing the overall population; specifically, β is a weighted average of the study-specific effects β_i seen in the individual study populations, where the weight for study population i is proportional to both the number of subjects selected (η_i) and to how much information is contributed per subject (ϕ_i). (We note this definition would also hold under the common-effect model, although there the choice of weights is redundant, as all β_i are assumed equal.)

The close connection between (3) and the estimator $\hat{\beta}$ is emphasized if we re-write (1) as

$$\hat{\beta} = \frac{\sum_{i=1}^{k} \frac{n_i}{\sum_{j=1}^{k} n_j} \frac{1}{n_i \hat{\sigma}_i^2} \hat{\beta}_i}{\sum_{i'=1}^{k} \frac{n_i'}{\sum_{j=1}^{k} n_j} \frac{1}{n_i' \hat{\sigma}_i'^2}}.$$

Comparing each term to its equivalent in (3), we see that $\frac{n_i}{\sum_{j=1}^{k} n_j}$ is consistent for η_i, $\frac{1}{n_i \hat{\sigma}_i^2}$ is consistent for ϕ_i, and $\hat{\beta}_i$ is consistent for β_i, making $\hat{\beta}$ consistent for β, under the fixed-effects model. As already noted, under the standard conditions (2) provides an appropriate standard error, thus validating use of $\hat{\beta}$ and $SE(\hat{\beta})$ under the fixed-effects model. Furthermore, under this model $\hat{\beta}$ has good statistical properties; under the standard conditions, the σ_i^2 and proportions η_i are known with negligible error, and so if each β_i is unbiased, so is $\hat{\beta}$. Furthermore, $\hat{\beta}$ is efficient among estimates that are unbiased for β.

These results directly contradict statements prevalent in both the statistical literature and in advice to non-statisticians, that $\hat{\beta}$ and $SE(\hat{\beta})$ are only valid when the β_i (or $\hat{\beta}_i$) are homogeneous. Under standard conditions, we have derived them for inference on β, a weighted average of the β_i. No assumption of homogeneity

was made, or is required. It remains to discuss the scientific relevance of β; the use of this specific weighted average is described in many common settings in Sections 3.1 and 3.2 below, with more general results given in Section 3.3. Analysis using other weighted averages in considered in Appendix C.

3.1 Inverse-variance weighted average in the linear regression context

We consider the relevance of the parameter β defined by (3), in situations where it may be possible to pool all the data and perform a comprehensive one-stage analysis, or 'individual participant data meta-analysis' (Higgins et al., 2001). Here, we assume that interest lies in the linear association between outcome Y and a single covariate X, and that the variance of Y is constant across all observations within all studies.

In this setting, a natural and common 'one-stage' analysis performs multiple linear regression of Y on X adjusting for study, i.e. fitting study-specific intercepts. As shown in Appendix A, this regression's X coefficient is consistent for β as defined in (3). Even though the one-stage analysis does not explicitly compute different estimates from each study, it estimates exactly the same quantity as combining study-specific estimates in a fixed-effects meta-analysis. This equivalence holds regardless of whether the true β_i are heterogeneous or not. Interpreting β through the one-stage approach is therefore just as valid as through meta-analysis, which may be helpful in practice for audiences more used to regression analyses than weighting or stratification.

For example, in Figure 1's genetic meta-analysis, we may interpret β as the average difference in blood pressure (Y) per additional copy of the 'A' allele (X) in subjects the same (unspecified) study population. This is a reasonable summary measure – and is the default in practice – but caveats familiar from linear regression also suggest its limitations.

First, β is an average effect, averaging over the whole population contributing to the analysis, and therefore may not capture the association in any subgroups (Berrington de González and Cox, 2007). Even assuming homogeneity, i.e. that the effects are identical in each study population, there is no guarantee that effects are identical within any of those study populations. (See e.g. Berman (1988) and Buja et al. (2016) for the interpretation of linear regression as an 'average slope'.)

Second, β averages difference in blood pressure per unit difference in the covariate of interest. In this example no distinction is made between having, say, 0 vs 1 or 1 vs 2 copies of the 'A' allele – nor is any assumption needed that the effects of these differences are identical. But consequently inference on β

alone says nothing about whether the mean outcome is linear in this covariate (an 'additive' effect) or follows 'dominant' or 'recessive' patterns (Brooker, 1999, Chapter 2).

3.2 Inverse-variance weighted average and the Mantel-Haenszel method

The behavior demonstrated in Section 3.1 is not limited to continuous outcomes, as we now illustrate by considering meta-analysis of data from 2×2 tables. We use standard notation for binary covariate X and binary outcome Y, with cell counts for each study i denoted as in Table 1. In this notation, the classical Mantel-Haenszel estimator (Mantel and Haenszel, 1959; Kuritz et al., 1988) is

$$\hat{\psi} = \frac{\sum_{i=1}^{k} \frac{A_i D_i}{n_i}}{\sum_{i'=1}^{k} \frac{B_{i'} C_{i'}}{n_{i'}}}.$$

It provides a measure of the observed association between the 'row' and 'column' variables, across a series of 2×2 tables. It can also be written in the form of equation (1), as

$$\begin{aligned} \hat{\psi} &= \frac{\sum_{i=1}^{k} \Omega_i \frac{A_i D_i}{B_i C_i}}{\sum_{i'=1}^{k} \Omega_{i'}}, \\ \text{where } \Omega_i &= \frac{n_i}{\sum_{i'=1}^{k} n_{i'}} \frac{B_i}{n_i} \frac{C_i}{n_i}. \end{aligned}$$

Written this way, we see that $\hat{\psi}$ is a weighted average of odds ratio estimates specific to each population i, i.e. a form of meta-analysis, and that it is consistent for

$$\psi = \frac{\sum_{i=1}^{k} \eta_i \phi'_i \psi_i}{\sum_{i'=1}^{k} \eta_i \phi'_i},$$

where ϕ'_i denotes the odds ratio of association in subpopulation i, η_i is again the proportion of the population that comes from study population i, and ϕ'_i, the product of the limiting values of B_i/n_i and C_i/n_i, the proportion of observations in each subpopulation that lie in each off-diagonal position.

As in Section 3.1, the definitions of ψ and $\hat{\psi}$ upweight populations where a greater proportion of the data would lie in the off-diagonal positions (i.e. greater B_i/n_i and C_i/n_i), and for larger populations (i.e. greater η_i); re-weighting the estimate to reflect other populations is straightforward (Greenland, 1982). As well as being robust to assumptions of homogeneity – a property not shared by competing methods such as conditional logistic regression – the Mantel-Haenszel estimate is generally efficient under homogeneity (Donner and Hauck, 1986; Hauck

		Outcome, Y		Total
		0	1	
Covariate, X	0	A_i	B_i	$A_i + B_i$
	1	C_i	D_i	$C_i + D_i$
Total		$A_i + C_i$	$B_i + D_i$	n_i

Table 1: Notation for study i, when combining multiple 2×2 tables

and Donner, 1988; Yanagimoto, 1990), and compared to other estimates has a more stable target of inference under heterogeneity (McKinlay, 1978). As a default method, giving a summary of the association across the population, it therefore has much to recommend it (Breslow, 1996).

In the absence of additional covariates, and under the standard conditions, Wald tests based on the Mantel-Haenszel estimate are asymptotically equivalent to those from logistic regression analysis of the row variable on the column variable which adjusts for study, i.e. fits study-specific intercepts, and are similarly asymptotically equivalent to those from conditional logistic regression, stratifying on study (Breslow, 1981). The equivalence is not just asymptotic; the Mantel-Haenszel test for zero average effect is in fact identical to the score test in the conditional logistic regression model, at any sample size (Day and Byar, 1979), and the Mantel-Haenszel effect estimate is a one-step approximation to the logistic regression odds ratio (Breslow, 1981). We see that, as in the linear regression setting, meta-analysis of data from 2×2 tables estimates the same parameter as a combined-data analysis.

3.3 Inverse-variance weighted average for general regression

The results in Sections 3.1 and 3.2 in fact hold, in large samples, in many regression settings including generalized linear models, the Cox proportional hazards model, and linear mixed models for longitudinal data. Formal proofs, regularity conditions and illustrative simulations, given by Lin and Zeng (2010), are beyond the scope of this paper, but their two main results are as follows;

First, the parameter estimated by a combined-data regression analysis, that adjusts for study, can also be estimated by fixed-effects meta-analysis, under the standard conditions. This follows because the parameter is a weighted average of the true effects in the separate studies, with weights proportional to the precision provided by each study's data. The equivalence holds even when each study adjusts for covariates, and even if the covariates differ between studies.

Second, without further assumptions, the meta-analysis and combined-data

estimates of the parameter have similar asymptotic efficiency, i.e. essentially no information about parameter β is lost by using meta-analysis instead of combined-data analysis. The asymptotic efficiency is identical under homogeneity, and similar under heterogeneity unless the values of β_i differ drastically. When stronger assumptions can be made, for example that any additional covariate effects are identical in all studies, there is potential for a gain in efficiency by combined-data analysis, albeit a gain which is small in most situations. In particular, when the studies are randomized trials the gain is zero, since the additional covariates will be independent of randomized treatment.

4 Fixed-effect(s) meta-analysis and the goals of the inference

In Section 3, we saw that without any assumptions of homogeneity, the fixed-effects meta-analysis estimates a well-defined population parameter, of general relevance in realistic settings. Consequently, assessing the appropriateness of fixed-effects analyses by checking homogeneity is without foundation — yet exactly this argument appears throughout the literature, resulting in the difficulties described in Section 1.1. Both in theory and in practice, the argument is not tenable and should not be made.

However, if we erase this argument from the statistical lexicon, the question remains of how to decide whether to use fixed-effects meta-analysis. Of course, there is no single universally-correct answer, but to aid meta-analysts as they address this question for their analyses, in Sections 4.1, 4.2 and 4.3 we consider inferential goals in three practical situations: testing, estimation and prediction.

4.1 Testing

First we consider testing, and in particular tests of the so-called 'strong' null hypothesis, that all parameters β_i are exactly zero. This test would be of interest, say, if an exposure was hypothesized to have no effect in any of the populations under study, for example in clinical trials of a new drug treatment. Under the strong null hypothesis, the common-effect and fixed-effects models are both valid. Furthermore, we might note that a special case of the random-effects model also holds, where both the underlying mean and the between-study heterogeneity are zero. Under the strong null, it follows straightforwardly that parameter β from equation (3) is zero, and hence that Wald tests using $\hat{\beta}$ and $SE(\hat{\beta})$ from the fixed-effects analysis are appropriate, under the standard conditions (Senn, 2007). Moreover, due to the efficiency of $\hat{\beta}$, among linear estimates, these particular tests will have good power compared to a wide range of competitors.

A second testing-based inference asks "is the effect size 'on average' different from zero, in the study populations at hand?" The weighted average we use in (3) specifies precisely what 'average' we mean, and the 'weak' null hypothesis being tested is that this β is zero – which happens under the strong null, but also in some situations where individual β_i are non-zero. Under the standard conditions, valid tests of the weak null are given by Wald tests that compare $Z = \hat{\beta}/SE(\hat{\beta})$ to a standard Normal distribution. The genetic setting of Figure 1 provides a practical example where tests are commonly used in this way, whether or not heterogeneity is observed (Gögele et al., 2012). Not only is power to reject the strong null relevant, but the average effect size (in particular its direction) is used in planning replication analyses, performed in similar populations where similar average effects are plausible.

In settings where the weak null's $\beta = 0$ hypothesis is not of interest, and if the strong null that all $\beta_i \equiv 0$ is known not to hold, then tests based on inverse-variance weighted averages are not appropriate, under either the common-effect or fixed-effects model. It is possible that tests based on random-effects analyses would still be appropriate, assessing the weak null where the distribution of effect sizes has mean zero — but this subtle distinction between different weak nulls seems seldom likely to be useful in practice.

4.2 Estimation

When the goal is estimation, the relevance of common-effect and fixed-effects models will depend on the nature of the studies in the meta-analysis. Common-effect models can be justified only when the studies are near replicates of each other, so that is it reasonable to assume an identical parameter underlying every study. This is possible, for example when repeating carefully-controlled experiments, where the impact of e.g. 'batch effects' can be ignored, but is rare in less well-controlled forms of study.

When between-study differences are believed to exist, the fixed-effects parameter β as defined in (3) may be a useful parameter to estimate. For example, if the subjects in the studies contributing to the meta-analysis are representative of an overall population of interest, the fixed-effects estimate is directly relevant to that overall population — as would be the case in Figure 1. If, however, the sample sizes across studies vary so greatly that the combined population is unrepresentative of any plausible overall population, then the fixed-effects parameter will not be as useful. For example, clinical trials of the effect of intravenous magnesium for acute myocardial infarction include several small studies and two very large studies (the ISIS-4 trial randomized 58,050 participants and the MAGIC trial randomized 6213; see Li et al. (2007) for details). The observed effects in smaller trials indicated a benefit of magnesium, in contrast to the strongly null findings of

the two large trials. The usefulness of the fixed-effects parameter then depends on a judgment about whether the combined population across trials represents the population of interest, or whether populations under study in the smaller trials were under-represented and therefore deserve to be 'upweighted' in the analysis.

The natural estimation in random-effects analysis is of the mean of the random-effects distribution; when the samples are considered a simple-random sample from a population of possible studies, this mean represents the effect size we would expect to see in a new study, sampled in the same way (Higgins et al., 2009). However, if the random-effects assumption is motivated through exchangeability alone, there is no direct motivation of this mean parameter as the target of inference, beyond computational convenience, and further clarification should be sought regarding what constitutes a good or bad estimate.

4.3 Prediction

Using summary statistics $\hat{\beta}_i$ and $\hat{\sigma}_i$ alone, it is generally not possible to make well-calibrated predictions of new subject-level data, under any of the models discussed here. However, for some analyses we may want use the existing data to predict the parameter value a new study might estimate. For example, to estimate the power of a proposed new study, we would estimate the effect size of a new β_i^*. Estimates of the generalizability of a meta-analysis' findings would use typical values of new β_i^*, but also their spread.

In order to make these forms of inference, we must assume some connection between new β_i^* and the β_i studied in the original meta-analysis. Under the common-effect model, where by assumption the same parameter governs association in all studies, the connection is technically trivial; estimate $\hat{\beta}$ and $SE(\hat{\beta})$ supply everything we need to predict or generalize. Conversely the assumptions of the fixed-effects model, by themselves, say nothing about studies except those in the meta-analysis, and so offer no obvious means of predicting or generalizing the results. Additional assumptions could be added, about how different new effects $\hat{\beta}_i^*$ might be to the β_i in the original analysis, but the arbitrariness of doing so is a notable limitation. Between these two extremes lies the random-effects model, where under the assumption of effects being drawn randomly from a mixing distribution, we can naturally model future effects $\hat{\beta}_i^*$ as new draws from a population, the parameters of which are informed by the original $\hat{\beta}_i$ and $\hat{\sigma}_i$. For examples and further discussion see Higgins et al. (2009).

5 Dealing with heterogeneity alongside a fixed-effects meta-analysis

In Section 4, we described how fixed-effects analysis may be appropriate with or without assumptions that all effect sizes β_i are identical. However, evaluating the similarity of the β_i may still be desirable, when there is interest in parameters beyond the precision-weighted average effect β. For example, in the setting of Figure 1, while the contributing populations are broadly similar (e.g. all of European ancestry) they also differ by design in e.g. average age, proportion male/female, and other characteristics, and any genetic effect of the 'A' allele may differ between populations because of this. Testing homogeneity of the β_i allows us to detect these differences, and thus to reject the null hypothesis that the 'A' allele has zero effect in all populations; implementation of such tests is described in Section 5.1. Methods that quantify how the effect sizes β_i differ with regard to particular population characteristics are described in Section 5.2. All these methods are derived, as before, under the fixed-effects model, and the inference they provide is a complement to the fixed-effects analysis seen in Section 3.

5.1 Fixed-effects tests of homogeneity

Cochran's Q statistic is a well-known measure of heterogeneity in meta-analyses. In our notation, it is defined as

$$Q = \sum_{i=1}^{k} w_i(\hat{\beta}_i - \hat{\beta})^2.$$

We see that Q sums the squared deviation of each $\hat{\beta}_i$ from $\hat{\beta}$, weighting these summands by w_i. This choice of weights can be motivated informally; any contribution to Q that tells us more about β should also tell us more about how the β_i vary around β, and hence we should give it more weight. The form of weights can also be motivated by noting that it results in a statistic that is invariant to scale and location transformations of the $\hat{\beta}_i$. A less well-known way to motivate Q is by writing it as

$$Q = \left(\sum_{i=1}^{k} Z_i^2\right) - Z^2, \tag{4}$$

where $Z_i = \hat{\beta}_i/\hat{\sigma}_i$ and $Z = \hat{\beta}/SE(\hat{\beta})$. Here we obtain Q by subtraction, taking away the test statistic for assessing the 'weak' null from the test statistic assessing the strong null. Equivalently, we see that Q is measuring how significantly the study effects differ from zero, over and above how significantly their overall average differs from zero.

Several distributional properties of Q are known, all of which are stated here under the standard conditions. As shown in Appendix B, Q and Z are independent. Under the null hypothesis of homogeneity, including the 'strong' null of no effect in any study, Q has a χ^2 distribution with $k-1$ degrees of freedom; comparing Q to this reference distribution gives a well-known test of homogeneity. The Q statistic can be used as the basis for other ways to quantify similarity of effects across studies, for example using the I^2 statistic (Higgins and Thompson, 2002) or a straightforward method-of-moments estimate of the between-study variance (DerSimonian and Laird, 1986). Under heterogeneity Q has a non-central χ^2_{k-1} distribution with non-centrality parameter $\sum_{i=1}^{k} w_i(\beta_i - \beta)^2$ (Hedges and Pigott, 2001); in many practical settings this value will be small (Higgins et al., 2003) so the power of the test will be small. For more detailed discussion of this test see Hoaglin (2016). The distribution of Q when the standard conditions do not hold is explored by Kulinskaya et al. (2011a,b).

For our discussion, independence of Q and Z is the most important property. It re-emphasizes that testing for homogeneity and testing for no overall main effect are distinct goals, which may or may not be relevant to any specific meta-analysis. When testing is appropriate, this independence property also makes it straightforward to control Type I error rates, although what constitutes a Type I error will depend on one's scientific goals, and deserves careful prior consideration. To emphasize the variety of possible goals, we consider when six distinct testing approaches would be appropriate;

- *Test only the weak null, of zero average effect, in a fixed effects analysis.* As discussed in Section 3, this would be appropriate if the meta-analysis substitutes for combined analysis of the original datasets.

- *Perform only the test of homogeneity.* This would be appropriate if the presence or absence of an overall main effect was uninteresting. It would be a reasonable first step if, say, an established treatment with a known overall main effect was to be assessed for efficacy across a range of subpopulations.

- *Perform tests of the weak null and of homogeneity.* This would be appropriate if *any* deviation from the strong null is of interest and if we additionally want to distinguish between the alternatives of an overall main effect or homogeneity of effect between the subpopulations. To preserve the Type I error rate across both tests, significance levels of the individual tests would have to be adjusted, relative to their use as stand-alone tests. However, as the individual tests are independent, Bonferroni correction makes this adjustment simple, without unduly conservative inference. Other corrections could be used if we wish to 'spend' more of the Type I error rate on the test of overall main effect, rather than on the homogeneity test.

- *Follow a rejection of the weak null with a test of homogeneity.* This would be appropriate if our primary interest is the presence/absence of an overall main effect, but we secondarily want to assess homogeneity of effect across subpopulations. This would be reasonable if the subpopulations had some *a priori* distinction, such as sex or age, which was plausibly relevant to the effect size. (See also Section 5.2.) To preserve the Type I error across both tests, significance levels would have to be adjusted in some way (Siegmund, 1985), but with only two independent tests this is typically straightforward.

- *Perform a single test of the strong null.* This would be appropriate if any deviation from the strong null is of interest, but where we do not want to distinguish between the alternatives of an overall main effect and/or homogeneity of effect between the subpopulations. The test could be performed using the individual tests' results alone (as when we perform both tests), or by comparing a test statistic of the form $(1-\lambda)Q + \lambda Z^2$ to the convolution of $(1-\lambda)\chi^2_{k-1}$ and $\lambda\chi^2_1$, for some pre-specified constant λ. The values of λ that lead to good power will depend on plausible alternative hypotheses, i.e. good choices of λ will be context-specific. However, if it is plausible that the β_i are all similar, setting $\lambda = 1$ and testing the strong null using Z^2 alone will provide close to optimal power.

- *Do no tests.* Testing hypotheses need not be part of statistical inference. If we have no interest in whether the true values of β_i or β can be distinguished from pre-specified null values, or expect to have little power to make such distinctions, then motivation for statistical testing is weak. Many recent texts on meta-analysis have been particularly dismissive of testing for heterogeneity, instead favoring attempts to quantify heterogeneity (See Section 4.2). More straightforwardly, if we wish to know the range of parameter values that are plausible in particular populations, then confidence intervals provide more useful inference than tests (Borenstein et al., 2009; Higgins et al., 2003).

Whenever heterogeneity is of interest, it is natural and reasonable to examine which studies' effects are furthest from the others. Measuring distance in terms of statistical significance, the studies may be ordered by their contributions $w_i(\hat{\beta}_i - \hat{\beta})^2$ (Thompson, 1993) although the random variation in this observed ordering may be substantial. As shown in Figures 1 'forest plots' of the $\hat{\beta}_i$ and corresponding confidence intervals are a less formal approach that can identify outlying study results; for further plotting recommendations see Anzures-Cabrera and Higgins (2010).

5.2 Fixed-effects meta-regression

In Section 5.1, we addressed the presence of heterogeneity, as a secondary step in a fixed-effects analysis. When study-specific covariates are available, we can also extend the fixed-effects analysis to describe how the underlying effects β_i differ, among studies with different covariates; *meta-regression* is the general term for methods that do this. We note this option is not available under common-effect assumptions, as these have no heterogeneity to explain.

Figure 3 illustrates a linear meta-regression of the data in Figure 1, where we assess how the genetic effect of the 'A' allele differs among study populations with different mean ages. We see that, based on this data, the genetic effect appears not to be uniform but trends upwards as we consider studies with older average age – and in studies of complex disease outcomes like blood pressure this form of 'gene-environment interaction' is highly biologically plausible (Caspi and Moffitt, 2006), even in situations where the overall average effect is close to zero. We see that the size of the observed trend can be a useful summary of available data.

To formally justify the analysis in Figure 3 under the fixed-effects model, we assume (for simplicity) that a single univariate covariate x_i is available, known with negligible error in each study population. In a fixed-effects linear meta-regression, using the same notation as before, we estimate the linear trend by

$$\begin{aligned} \hat{\beta}_{MR} &= \frac{\sum_{i=1}^{k} w_i(\hat{\beta}_i - \hat{\beta})(x_i - \hat{x})}{\sum_{i'=1}^{k} w_{i'}(x_{i'} - \hat{x})^2}. \\ \text{where } \hat{x} &= \frac{\sum_{i=1}^{k} w_i x_i}{\sum_{i'=1}^{k} w_{i'}}. \end{aligned}$$

One way to motivate $\hat{\beta}_{MR}$ is through weighted linear regression of the $\hat{\beta}_i$ on the x_i (Van Houwelingen et al., 2002), with weights w_i. In doing so, derivations typically assume that the underlying β_i are truly linear in x_i – implying that any two studies with the same x_i would have identical β_i. However, much like the assumption of homogeneity in Section 3 this is not actually required. Following Berman (1988), we can instead view $\hat{\beta}_{MR}$ as just a weighted average of observed 'pairwise slopes' of the form $\frac{\hat{\beta}_i - \hat{\beta}_j}{x_i - x_j}$, for $i \neq j$. The weight is larger for studies that are more precise, and also where the difference between x_i and x_j is greater.

As with the discussion of fixed-effects estimate $\hat{\beta}$ in Section 3, $\hat{\beta}_{MR}$ estimates a population parameter; this parameter can be formally defined as

$$\begin{aligned} \beta_{MR} &= \frac{\sum_{i=1}^{k} \eta_i \phi_i (\beta_i - \beta)(x_i - x)}{\sum_{i'=1}^{k} \eta_{i'} \phi_{i'} (x_{i'} - x)^2} = \frac{\sum_{i=1}^{k} \eta_i \phi_i^2 (x_i - x)^2 \frac{\beta_i - \beta}{x_i - x}}{\sum_{i'=1}^{k} \eta_{i'} \phi_{i'} (x_{i'} - x)^2}, \qquad (5) \\ \text{where } x &= \frac{\sum_{i=1}^{k} \eta_i \phi_i x_i}{\sum_{i'=1}^{k} \eta_{i'} \phi_i}. \end{aligned}$$

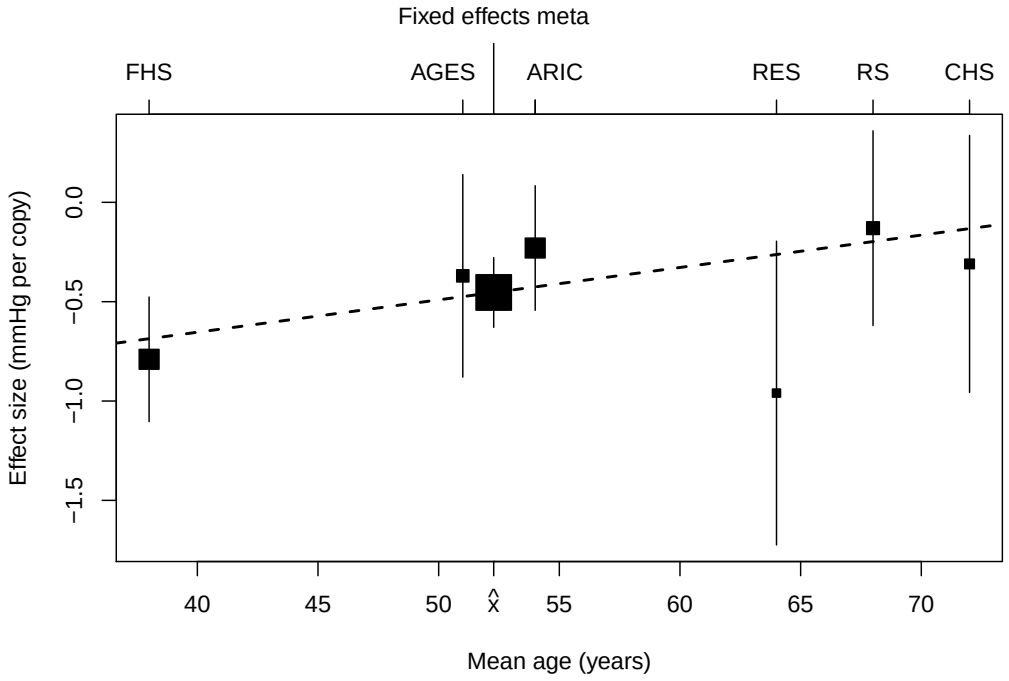


Figure 3: Meta regression, using data from Levy et al. (2009). The individual study effects from Figure 1 are plotted against their corresponding average age (x_i). The overall main effect is plotted at $\hat{x}$, the weighted average mean age. Box sizes are proportional to the precision of corresponding effect estimates. The dashed line indicates the fitted slope $\hat{\beta}_{MR}$ from meta-regression of the β_i on x_i.

The parameter β_{MR} is a slope, describing an average difference in β_i (measured relative to β) per unit differences in x_i (measured relative to x). The average is weighted; as with overall effect β we upweight contributions with greater η_i and greater ϕ_i, but here contributions with more extreme x_i are also upweighted – just as in simple linear regression.

To describe the uncertainty in $\hat{\beta}_{MR}$, the usual estimate of the standard error is obtained through weighted linear regression, and is

$$SE\left(\hat{\beta}_{MR}\right) = \sqrt{\frac{1}{\sum_{i=1}^{k} w_i (x_i - \hat{x})^2}}. \tag{6}$$

When true linearity of the β_i with x_i does not hold, this is not in general a valid estimate of the variability of $\hat{\beta}_{MR}$, and statistical inference based on it will not be valid (Higgins and Thompson, 2004). However, use of a heteroskedasticity-consistent or 'robust' estimate of standard error corrects this problem (Huber, 1967; White, 1980; Hedges et al., 2010).

We emphasize that in practice the estimation of this trend and its standard error will often be sufficient for inference based on the available data. However, if hypothesis tests are required, we can define

$$Z_{MR} = \frac{\hat{\beta}_{MR}}{SE\left(\hat{\beta}_{MR}\right)} = \frac{\sum_{i=1}^{k} w_i (\hat{\beta}_i - \hat{\beta})(x_i - \hat{x})}{\sqrt{\sum_{i'=1}^{k} w_{i'} (x_{i'} - \hat{x})^2}},$$

a natural test statistic for the null hypothesis that average study effect, averaging over studies with the same value of x_i, is the same at all levels of x_i – or less formally, that there is no trend of the β_i with x_i. This is a weaker null hypothesis than stating that all the β_i are zero, i.e. the 'strong' null, and in fact the test statistic for the strong null can be decomposed as

$$\begin{aligned} \sum_{i=1}^{k} Z_i^2 &= Z^2 + Z_{MR}^2 + Q_{MR}, \qquad (7) \\ \text{where} \quad Q_{MR} &= \sum_{i=1}^{k} w_i \left(\hat{\beta}_i - \hat{\beta} - \hat{\beta}_{MR}(x_i - \hat{x})\right)^2. \end{aligned}$$

This decomposition, in which the three terms are independent (see Appendix B) extends that seen in (4). Here, we see that Q_{MR} quantifies the significance of the variability in the β_i around zero that is not explained by the average effect (Z^2) or the trend with the x_i (Z_{MR}^2). Assuming that the β_i are truly linear in the x_i, and under the standard conditions, Z_{MR} has a $N(0,1)$ distribution, and Q_{MR} has a χ^2_{k-2} distribution.

As with the discussion of average effect β in Section 3.1, it makes sense to ask whether the linear trend β_{MR} is of substantive scientific interest. Linearity *must* hold under the strong null, when all β_i are zero, and so for tests of this null it seems uncontroversial that the answer has to be 'Yes'. In the presence of a truly non-linear but increasing trend, β_{MR} will usually capture most of variability in the β_i around their overall mean β, but use of one-dimensional summary measure $\hat{\beta}_{MR}$ may be inefficient compared to more flexible approaches. An extreme example of inefficiency comes when a symmetric U-shaped relationship truly holds between β_i and x_i, where the zero slope we estimate using meta-regression is not a helpful summary, and does not provide power to detect heterogeneity. In such settings the more general goal of estimating the 'response surface' may be appropriate, i.e. characterizing how β_i differs between many different populations (Rubin, 1992).

The results above follow from well-known properties of 'simple' linear regression, involving only one covariate. While not pursued here, analogous results hold for multiple regression, with more covariates. (For estimation of high-dimensional response surfaces see Ritz et al. (2008).) A more fundamental point is that meta-regression can be viewed as a natural extension of the fixed-effects analysis discussed in Section 2, which in many applications will be the primary focus. Meta-regression can be seen as a natural way to explore heterogeneity of the β_i around β, orthogonal to the inference on β itself, but using the same fixed-effects assumptions specified in Section 2.2.

One important difference between meta-regression and estimation of an overall main-effect is that the meta-regression does not, in general, estimate the same parameter we would estimate if individual level data were available. Due to the well-known ecological bias, the relationship between studies' point estimates and their study-wide average covariate level x_i need not reflect the relationship between individual outcomes and covariates (Berlin et al., 2002). The differences can be stark; sign-reversals can be plausible (Teramukai et al., 2004). While no meta-regression method that uses study-specific average covariates can eliminate all risk of ecological bias (a form of confounding) the problem can be reduced by introducing measures of within-study variability of covariate x_i into the analysis. An important special case when this is not required is testing the strong null, of no association between outcome and covariate, or between $\hat{\beta}_i$ and x_i.

6 Discussion and Limitations

Fixed-effects meta-analysis is a fundamental analytic tool for evidence synthesis, and confusion about it is therefore a fundamental obstacle to successful evidence synthesis. As we have described, the fixed-effect(s) analysis has various interpretations, but can be simply and robustly motivated via consideration of an overall

average effect parameter, where the weighting in this average reflects what an analyst might use were it possible to combine individual-level data.

The average effect discussed here (i.e. parameter β defined in Equation (3)) can be estimated using straightforward methods, which have good statistical properties under the standard conditions. These methods extend naturally, and without restrictive modeling assumptions, to descriptions of variability around the overall average effect (i.e. Q as defined in Equation (4)) and to trends in effects seen across studies (i.e. β_{MR} as defined in Equation (5)).

When the standard conditions do not hold, caution about the validity of these methods and their applied results is merited. One plausible violation occurs when combining results from small studies, when the statistical imprecision in the study's estimated standard errors is not negligible. In this situation, the estimates and test statistics described here will typically be heavier-tailed than the approximations given here. Subject to more assumptions, better approximations to the true distributions are available in the literature (Brazzale et al., 2007; Brazzale, 2005) but are not yet fully-integrated into software dedicated to meta-analysis. With small studies, a similar concern is that the study estimates are not close to Normal, particularly when obtained from non-linear forms of regression, or other more complex methods. In both situations, default frequentist inferences may be poorly calibrated; see e.g. Hoaglin (2015) for a recent discussion. Furthermore, for some effect measures (including log odds ratios) the statistical correlation between estimated standard errors and estimated effect sizes is particularly problematic in small studies (Shuster, 2010). In some relatively simple setting where only statistical tests are required, all of these small-sample issues can be addressed by permutation methods, see e.g. Hong et al. (2006). For other goals, bootstrap, parametric bootstrap and more sophisticated likelihood methods may be helpful (Van Den Noortgate and Onghena, 2005; Guolo, 2012), and realistic simulation studies may help the user identify which methods plausibly work well for their situation (Hoaglin, 2015).

Another important violation of the standard conditions occurs when studies do not provide valid analyses, either because of limitations in the design and conduct of the study, or because, after data collection, *post hoc* changes are made to the analysis, but reported analyses do not take these steps into account. In a single analysis such procedures may bias effect estimates, and seriously underestimate standard errors (Leeb and Pötscher, 2005); they can similarly invalidate inference from any meta-analysis. If in practice these procedures cannot be avoided, accounting for the biases they induce is known to be difficult, not least because the biases are difficult to identify in the first place.

A final consideration is less directly statistical; is the population represented by the amalgamation of the k contributing studies of interest? Analytic techniques for re-weighting may adjust the standard fixed-effects analysis accord-

ingly, but the same concerns are addressed when we select which studies (and thus which study populations) will be included in the meta-analysis, or excluded. This step is stressed by, for example, Higgins and Green (2011, Chap 5,6,7), but inherently offers each study's data an all-or-nothing option. When a re-weighting scheme can be established, we may instead retain some contribution from studies which would otherwise be excluded; further discussion is given in Appendix C.

To conclude, when synthesizing evidence on a univariate effect, fixed-effects meta-analysis is a simple method that provides useful well-calibrated inference about an average effect, under mild conditions. Fixed-effects meta-analysis can and often should be used in situations where effects differ between studies. This recommendation in no way rules out statistical analyses of heterogeneity, and fixed-effects meta-analysis has natural complements (Cochran's Q, meta-regression) that provide them. Measures of heterogeneity should not be used to determine whether fixed-effects analysis is appropriate, but users should instead make this decision by deciding whether fixed-effects analysis — or some variant of it — answers a question that is relevant to the scientific situation at hand.

References

Ades, A., N. Welton, D. Caldwell, M. Price, A. Goubar, and G. Lu (2008). Multiparameter evidence synthesis in epidemiology and medical decision-making. *Journal of Health Services Research & Policy 13*(suppl 3), 12–22.

Anzures-Cabrera, J. and J. Higgins (2010). Graphical displays for meta-analysis: An overview with suggestions for practice. *Research Synthesis Methods 1*(1), 66–80.

Berkey, C., D. Hoaglin, F. Mosteller, and G. Colditz (1995). A random-effects regression model for meta-analysis. *Statistics in Medicine 14*(4), 395–411.

Berlin, J., J. Santanna, C. Schmid, L. Szczech, and H. Feldman (2002). Individual patient-versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. *Statistics in Medicine 21*(3), 371–387.

Berman, M. (1988). A theorem of Jacobi and its generalization. *Biometrika 75*(4), 779–783.

Berman, N. and R. Parker (2002). Meta-analysis: neither quick nor easy. *BMC Medical Research Methodology 2*(1), 10.

Bernardo, J. and A. Smith (2003). *Bayesian Theory*. Wiley.

Berrington de González, A. and D. Cox (2007). Interpretation of interaction: A review. *The Annals of Applied Statistics 1*(2), 371–385.

Borenstein, M., L. Hedges, J. Higgins, and H. Rothstein (2009). *Introduction to Meta-Analysis*. Wiley Online Library.

Brazzale, A. (2005). hoa: An R package bundle for higher order likelihood inference. *R News 5*(1), 20–27.

Brazzale, A., A. Davison, and N. Reid (2007). *Applied Asymptotics: Case Studies in Small-sample Statistics*. Cambridge: Cambridge University Press.

Breslow, N. (1981). Odds ratio estimators when the data are sparse. *Biometrika 68*(1), 73–84.

Breslow, N. (1996). Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association*, 14–28.

Brooker, R. (1999). *Genetics: analysis and principles*. Addison-Wesley.

Buja, A., R. Berk, L. Brown, E. George, E. Pitkin, M. Traskin, L. Zhao, and K. Zhang (in press, 2016). Models as approximationsa conspiracy of random regressors and model deviations against classical inference in regression. *Statistical Science*.

Casella, G. and R. Berger (2001). Statistical inference.

Caspi, A. and T. Moffitt (2006). Gene–environment interactions in psychiatry: joining forces with neuroscience. *Nature Reviews Neuroscience 7*(7), 583–590.

Day, N. and D. Byar (1979). Testing hypotheses in case-control studies-equivalence of mantel-haenszel statistics and logit score tests. *Biometrics*, 623–630.

de Bakker, P., M. Ferreira, X. Jia, B. Neale, S. Raychaudhuri, and B. Voight (2008). Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Human molecular genetics 17*(R2), R122–R128.

DerSimonian, R. and N. Laird (1986). Meta-analysis in clinical trials. *Controlled clinical trials 7*(3), 177–188.

Donner, A. and W. Hauck (1986). The large-sample relative efficiency of the mantel-haenszel estimator in the fixed-strata case. *Biometrics*, 537–545.

Early Breast Cancer Trialists' Collaborative Group (1998). Tamoxifen for early breast cancer: an overview of the randomised trials. *Lancet 351*, 1451–1467.

Fleiss, J. (1993). Review papers: The statistical basis of meta-analysis. *Statistical Methods in Medical Research 2*(2), 121–145.

Gögele, M., C. Minelli, A. Thakkinstian, A. Yurkiewich, C. Pattaro, P. Pramstaller, J. Little, J. Attia, and J. Thompson (2012). Methods for meta-analyses of genome-wide association studies: Critical assessment of empirical evidence. *American Journal of Epidemiology 175*(8), 739–749.

Greenland, S. (1982). Interpretation and estimation of summary ratios under heterogeneity. *Statistics in Medicine 1*(3), 217–227.

Guolo, A. (2012). Higher-order likelihood inference in meta-analysis and meta-regression. *Statistics in Medicine 31*(4), 313–327.

Hartung, J., G. Knapp, and B. Sinha (2011). *Statistical meta-analysis with applications*, Volume 738. Wiley-Interscience.

Hauck, W. and A. Donner (1988). The asymptotic relative efficiency of the mantel-haenszel estimator in the increasing-number-of-strata case. *Biometrics*, 379–384.

Hedges, L. and T. Pigott (2001). The power of statistical tests in meta-analysis. *Psychological Methods 6*(3), 203.

Hedges, L., E. Tipton, and M. Johnson (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research synthesis methods 1*(1), 39–65.

Hedges, L. and J. Vevea (1998). Fixed-and random-effects models in meta-analysis. *Psychological methods 3*(4), 486.

Higgins, J. and S. Green (Eds.) (2011). *Cochrane Handbook for Systematic Reviews of Interventions*. The Cochrane Collaboration. Version 5.1.0 [updated March 2011], available from `www.cochrane-handbook.org`.

Higgins, J. and S. Thompson (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine 21*(11), 1539–1558.

Higgins, J. and S. Thompson (2004). Controlling the risk of spurious findings from meta-regression. *Statistics in Medicine 23*(11), 1663–1682.

Higgins, J., S. Thompson, J. Deeks, and D. Altman (2003). Measuring inconsistency in meta-analyses. *BMJ 327*(7414), 557.

Higgins, J., S. Thompson, and D. Spiegelhalter (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 172*(1), 137–159.

Higgins, J., A. Whitehead, R. Turner, R. Omar, and S. Thompson (2001). Meta-analysis of continuous outcome data from individual patients. *Statistics in Medicine 20*(15), 2219–2241.

Hoaglin, D. C. (2015). We know less than we should about methods of meta-analysis. *Research synthesis methods 6*(3), 287–289.

Hoaglin, D. C. (2016). Misunderstandings about Q and Cochran's Q test' in meta-analysis. *Statistics in Medicine 35*(4), 485–495.

Hong, F., R. Breitling, C. W. McEntee, B. S. Wittner, J. L. Nemhauser, and J. Chory (2006). Rankprod: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics 22*(22), 2825–2827.

Huber, P. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Volume 1, pp. 221–33.

Hunter, J. and F. Schmidt (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage Publications, Inc.

Konstantopoulos, S. and L. Hedges (2009). Analyzing effect sizes: Fixed-effects models. *The handbook of research synthesis and meta-analysis*, 279–293.

Kulinskaya, E., M. B. Dollinger, and K. Bjørkestøl (2011a). On the moments of cochran's q statistic under the null hypothesis, with application to the meta-analysis of risk difference. *Research synthesis methods 2*(4), 254–270.

Kulinskaya, E., M. B. Dollinger, and K. Bjørkestøl (2011b). Testing for homogeneity in meta-analysis i. the one-parameter case: Standardized mean difference. *Biometrics 67*(1), 203–212.

Kuritz, S., J. Landis, and G. Koch (1988). A general overview of mantel-haenszel methods: applications and recent developments. *Annual Review of Public Health 9*(1), 123–160.

Laird, N., S. Fitzmaurice, and X. Ding (2010). Comments on empirical vs natural weighting in random effects meta-analysis. *Statistics in Medicine 29*(12), 1266–1267.

Leeb, H. and B. Pötscher (2005). Model selection and inference: Facts and fiction. *Econometric Theory 21*(1), 21–59.

Leeflang, M., J. Deeks, C. Gatsonis, P. Bossuyt, et al. (2008). Systematic reviews of diagnostic test accuracy. *Annals of Internal Medicine 149*(12), 889–897.

Levy, D., G. Ehret, K. Rice, G. Verwoert, L. Launer, A. Dehghan, N. Glazer, A. Morrison, A. Johnson, T. Aspelund, et al. (2009). Genome-wide association study of blood pressure and hypertension. *Nature Genetics 41*(6), 677–687.

Li, J., Q. Zhang, M. Zhang, and M. Egger (2007). Intravenous magnesium for acute myocardial infarction. Cochrane Database of Systematic Reviews. Issue 2. Art. No.: CD002755. DOI: 10.1002/14651858.CD002755.pub2.

Lin, D. and D. Zeng (2010). On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika 97*(2), 321–332.

Lipsey, M. and D. Wilson (2001). *Practical meta-analysis*. Applied social research methods series. Sage Publications.

Lu, G. and A. Ades (2004). Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine 23*(20), 3105–3124.

Lumley, T. (2010). *Complex surveys: a guide to analysis using R*, Volume 565. Wiley.

Mantel, N. and W. Haenszel (1959). Statical aspects of the analysis of data from retrospective studies of disease. *Journal of the National cancer Institute 22*(4).

McKinlay, S. (1978). The effect of nonzero second-order interaction on combined estimators of the odds ratio. *Biometrika 65*(1), 191–202.

Olkin, I. and A. Sampson (1998). Comparison of meta-analysis versus analysis of variance of individual patient data. *Biometrics*, 317–322.

Pereira, T., N. Patsopoulos, G. Salanti, and J. Ioannidis (2009). Discovery properties of genome-wide association signals from cumulatively combined data sets. *American Journal of Epidemiology 170*(10), 1197–1206.

Peto, R. (1987). Why do we need systematic overviews of randomized trials? [with discussion]. *Statistics in Medicine 6*(3), 233–244.

Ritz, J., E. Demidenko, and D. Spiegelman (2008). Multivariate meta-analysis for data consortia, individual patient meta-analysis, and pooling projects. *Journal of Statistical Planning and Inference 138*(7), 1919–1933.

Rubin, D. (1992). Meta-analysis: Literature synthesis or effect-size surface estimation? *Journal of Educational and Behavioral Statistics 17*(4), 363–374.

Rücker, G., G. Schwarzer, J. Carpenter, and M. Schumacher (2010). Comments on empirical vs natural weighting in random effects meta-analysis by jj shuster, statistics in medicine 2009; 26, published online, doi: 10.1002/sim. 3607. *Statistics in Medicine 29*(28), 2963–2965.

Schervish, M. (1995). *Theory of Statistics*. New York: Springer.

Schmidt, F. and J. Hunter (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage Publications, Incorporated.

Senn, S. (2007). Trying to be precise about vagueness. *Statistics in Medicine 26*(7), 1417–1430.

Shuster, J. (2010). Empirical vs natural weighting in random effects meta-analysis. *Statistics in Medicine 29*(12), 1259–1265.

Siegmund, D. (1985). *Sequential analysis: tests and confidence intervals*. Springer.

Spiegelhalter, D., K. Abrams, and J. Myles (2004). *Bayesian approaches to clinical trials and health-care evaluation*, Volume 13. Wiley.

Stephens, M. and D. Balding (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics 10*(10), 681–690.

Teramukai, S., Y. Matsuyama, S. Mizuno, and J. Sakamoto (2004). Individual patient-level and study-level meta-analysis for investigating modifiers of treatment effect. *Japanese journal of clinical oncology 34*(12), 717–721.

Thompson, S. (1993). Controversies in meta-analysis: the case of the trials of serum cholesterol reduction. *Statistical Methods in Medical Research 2*(2), 173–192.

Thompson, S. and J. Higgins (2010). Comments on empirical vs natural weighting in random effects meta-analysis. *Statistics in Medicine 29*(12), 1270–1271.

Turner, R., D. Spiegelhalter, G. Smith, and S. Thompson (2009). Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 172*(1), 21–47.

Van Den Noortgate, W. and P. Onghena (2005). Parametric and nonparametric bootstrap methods for meta-analysis. *Behavior Research Methods 37*(1), 11–22.

Van Houwelingen, H., L. Arends, and T. Stijnen (2002). Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine 21*(4), 589–624.

Waksman, J. (2010). Comments on empirical vs natural weighting in random effects meta-analysis. *Statistics in Medicine 29*(12), 1268–1269.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, 817–838.

Whitehead, A. and J. Whitehead (1991). A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in Medicine 10*(11), 1665–1677.

Yanagimoto, T. (1990). Combining moment estimates of a parameter common through strata. *Journal of Statistical Planning and Inference 25*(2), 187–194.

Zhong, H. and R. Prentice (2010). Correcting winner's curse in odds ratios from genomewide association findings for major complex human diseases. *Genetic epidemiology 34*(1), 78–91.

7 Appendices

7.1 Appendix A: Estimation of fixed-effects β via one-stage analysis

To illustrate how the definition (3) relates to standard methods, we consider a comprehensive, one-stage analysis, or 'individual participant data meta-analysis' (Higgins et al., 2001); similar arguments are presented by Olkin and Sampson (1998). We assume that the $\hat{\beta}_i$ and $\hat{\sigma}_i$ are in each study obtained from simple linear regression analyses, from regression of length n_i vectors $\boldsymbol{Y}_i$ on $\boldsymbol{Z}_i$, where covariate z is measured on all subjects. Adjusting for study, the coefficient of z in a pooled linear regression is obtained as the last element of

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

where

$$\boldsymbol{X} = \begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} & \boldsymbol{z}_1 \\ \mathbf{0}_{n_1} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2} & \boldsymbol{z}_2 \\ \vdots & \vdots & \ddots & \vdots & \\ \mathbf{0}_{n_k} & \mathbf{0}_{n_k} & \cdots & \mathbf{1}_{n_k} & \boldsymbol{z}_k \end{bmatrix}$$

and $\boldsymbol{Y}$ denotes the length $\sum_{i=1}^{k} n_k$ vector of outcomes, i.e.

$$\boldsymbol{Y}^T = \{\boldsymbol{Y}_1^T, \boldsymbol{Y}_2^T, ..., \boldsymbol{Y}_k^T\}.$$

By the Frisch-Waugh-Lovell theorem, this coefficient can also be written as

$$\frac{Cov(\tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{Z}})}{Var(\tilde{\boldsymbol{Z}})}$$

where $Cov()$ and $Var()$ denote sample covariance and variance of vectors

$$\begin{aligned}
\tilde{\boldsymbol{Y}}^T &= \left\{\boldsymbol{Y}_1^T - \bar{\boldsymbol{Y}}_1 \mathbf{1}_{n_1}^T, \boldsymbol{Y}_2^T - \bar{\boldsymbol{Y}}_2 \mathbf{1}_{n_2}^T, \ldots, \boldsymbol{Y}_k^T - \bar{\boldsymbol{Y}}_k \mathbf{1}_{n_k}^T\right\} \\
\tilde{\boldsymbol{Z}}^T &= \left\{\boldsymbol{Z}_1^T - \bar{\boldsymbol{Z}}_1 \mathbf{1}_{n_1}^T, \boldsymbol{Z}_2^T - \bar{\boldsymbol{Z}}_2 \mathbf{1}_{n_2}^T, \ldots, \boldsymbol{Z}_k^T - \bar{\boldsymbol{Z}}_k \mathbf{1}_{n_k}^T\right\},
\end{aligned}$$

and where $\bar{\boldsymbol{Y}}_i = \boldsymbol{Y}_i^T \mathbf{1}_{n_i}/n_i$ and $\bar{\boldsymbol{Y}}_i = \boldsymbol{Y}_i^T \mathbf{1}_{n_i}/n_i$ denote the study-specific mean outcomes and mean covariates. The vectors $\tilde{\boldsymbol{Y}}$ and $\tilde{\boldsymbol{Z}}$ contain outcomes and covariates residuals, obtained after regressing out study-specific intercepts.

The coefficient of z is therefore

$$\begin{aligned}
\frac{\sum_i (\boldsymbol{Y}_i - \bar{\boldsymbol{Y}}_i \mathbf{1}_{n_i})^T (\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}_i \mathbf{1}_{n_i})}{\sum_{i'} (\boldsymbol{Z}_{i'} - \bar{\boldsymbol{Z}}_{i'} \mathbf{1}_{n_{i'}})^T (\boldsymbol{Z}_{i'} - \bar{\boldsymbol{Z}}_{i'} \mathbf{1}_{n_{i'}})} &= \frac{\sum_i \frac{(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}}_i \mathbf{1}_{n_i})^T (\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}_i \mathbf{1}_{n_i})}{(\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}_i \mathbf{1}_{n_i})^T (\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}_i \mathbf{1}_{n_i})} (\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}_i \mathbf{1}_{n_i})^T (\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}_i \mathbf{1}_{n_i})}{\sum_i (\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}_i \mathbf{1}_{n_i})^T (\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}_i \mathbf{1}_{n_i})} \\
&= \frac{\sum_i \hat{\beta}_i n_i Var(\boldsymbol{Z}_i)}{\sum_{i'} n_{i'} Var(\boldsymbol{Z}_{i'})} = \sum_i \hat{\beta}_i \frac{n_i Var(\boldsymbol{Z}_i)}{\sum_{i'} n_{i'} Var(\boldsymbol{Z}_{i'})}.
\end{aligned}$$

Under the assumption of constant variance of all the outcomes, we see that the parameter estimated using individual data is a weighted average of the β_i, where the weighting is proportional to $n_i Var(\boldsymbol{Z}_i)$, which is in turn proportional to $1/\sigma_i^2$ under the standard conditions. The same equivalences hold if we let each n_i tend to infinity (i.e. at the population level) while keeping their relative sizes constant – as is standard in large-sample arguments for single-study analyses. In this limit, regardless of homogeneity, the weighting is proportional to $\eta_i \phi_i$, and we see that the combined one-stage data analysis consistently estimates the same parameter of the same population as the fixed effects meta-analysis. To generalize the argument to include study-specific covariates, $\mathbf{1}_{n_i}$ can be replaced by the design matrix of covariates (excluding Z) for each study.

7.2 Appendix B: Properties of Z_{MR}, Q_{MR}

Notation used in the main body of the paper is assumed throughout. We note that, under the standard conditions, the σ_i are estimated by $\hat{\sigma}_i$ with negligible

error, and therefore develop arguments with σ_i known; in all cases, the same results hold asymptotically when the estimation of σ_i is taken into account.

The components of Q and Z can be written in block matrix form as follows;

$$\left[\begin{array}{c} \hat{\boldsymbol{\beta}} - \hat{\beta}\mathbf{1}_k \\ \hat{\beta} \end{array} \right] = \left[\begin{array}{c} I_k - \mathbf{1}_k \boldsymbol{w}^T/W \\ \boldsymbol{w}^T/W \end{array} \right] \hat{\boldsymbol{\beta}},$$

where $\hat{\boldsymbol{\beta}}$ denotes the length k vector of study-specific estimtes $\hat{\beta}_i$, $\mathbf{1}_k$ is a length k vector of ones, I_k denotes the $k \times k$ identity matrix, $\boldsymbol{w}$ is the length k vector of precisions $w_i, 1 \leq i \leq k$ and $W = \sum_{i=1}^{k} w_i$. Hence the relevant covariances are obtained from

$$\left[\begin{array}{c} I_k - \mathbf{1}_k \boldsymbol{w}^T/W \\ \boldsymbol{w}^T/W \end{array} \right] \Sigma \left[\begin{array}{cc} I_k - \boldsymbol{w}\mathbf{1}_k^T/W, & \boldsymbol{w}/W \end{array} \right]$$

where $\Sigma = \text{diag}\,(\{\sigma_i^2\})$, the covariance of $\hat{\boldsymbol{\beta}}$. The expression expands to

$$\left[\begin{array}{cc} \Sigma - \mathbf{1}_k \mathbf{1}_k^T/W & \mathbf{0}_k \\ \mathbf{0}_k^T & 1/W \end{array} \right]$$

where $\mathbf{0}_k$ denotes a length k vector of zeros.

In the meta-regression setting, we similarly seek the covariance of elements of

$$\left[\begin{array}{c} \hat{\boldsymbol{\beta}} - \hat{\beta}\mathbf{1}_k - \hat{\beta}_{MR}(\boldsymbol{x} - \hat{x}) \\ \hat{\beta} \\ \hat{\beta}_{MR} \end{array} \right],$$

where $\boldsymbol{x}$ denotes the length k vector with elements x_i. Hence we seek elements of

$$\left[\begin{array}{c} I_k - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W} \\ (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W} \end{array} \right] \Sigma \left[\begin{array}{cc} I_k - \boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T, & \boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1} \end{array} \right]$$

where $\boldsymbol{X}$ denotes the $k \times 2$ design matrix with columns $\mathbf{1}_k$ and $\boldsymbol{x}$, and where $\boldsymbol{W}$ denotes the $k \times k$ diagonal matrix with diagonal elements w_i. This expression simplifies to

$$\left[\begin{array}{cc} \Sigma - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T & \mathbf{0} \\ \mathbf{0} & (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1} \end{array} \right]$$

where the lower right hand entry is

$$\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} = \left[\begin{array}{cc} \sum w_i & 0 \\ 0 & \sum_i w_i(x_i - \hat{x})^2 \end{array} \right].$$

With zero covariance again implying independence (under the Normality implicit in the standard conditions) we find that the elements Q_{MR}, Z, and Z_{MR} are mutually independent.

7.3 Appendix C: Meta-analytic analogs of alternatively-weighted individual-data analyses

The focus of this manuscript is meta-analytic estimates (means, and regression slopes) that are defined via weighting. In our default meta-analytic approach, we have focused on weighting study-specific estimates so as to provide estimates of parameters in standard unweighted individual-data analyses. However, the same framework also supports interpreting meta-analytic approaches in terms of individual-data analyses that use non-uniform weights. Non-uniform weighting of this sort is standard in the statistical survey literature (Lumley, 2010), where close analogs of the following results can be found.

We assume the same basic setup as in Appendix A, except that we now consider weighted linear regression estimates of Y on Z, adjusting for study. For simplicity, we consider situations where each participant in study i receives the same weight, denoted ω_i. Using the same arguments as in Appendix A, the estimator from a one-stage analysis of individual data that adjusts for study is

$$\frac{\sum_i \hat{\beta}_i \omega_i n_i Var(\boldsymbol{Z}_i)}{\sum_i \omega_i n_i Var(\boldsymbol{Z}_i)},$$

which is consistent for

$$\frac{\sum_i \beta_i \omega_i \eta_i \phi_i}{\sum_i \omega_i \eta_i \phi_i}. \tag{8}$$

A simple first example of weights uses $1/\omega_i \propto \eta_i \phi_i$, giving an estimate consistent for the arithmetic mean $\sum_i \beta_i / k$. The meta-analytic estimator $\sum_i \hat{\beta}_i / k$ in consistent for the parameter defined in (8). This estimator has been suggested in practice (Shuster, 2010) – but as this argument shows, it corresponds to upweighting data from studies which represent a smaller proportion of the overall population (smaller η_i) or which have weaker designs (smaller ϕ_i). Hence beyond its simplicity the method is unlikely to be a good choice in practice, and it has attracted corresponding criticism (Laird et al., 2010; Rücker et al., 2010; Thompson and Higgins, 2010; Waksman, 2010).

A second example considers re-weighting individuals in order to reflect a particular target population. Specifically, we consider re-weighting the individuals contributing to the example of Figure 1 to reflect a population with some specified ratio of men to women, in which we want to estimate association between blood pressure and rs11014166 genotype. Measures of association within each study, and each study's sex ratio are given in Figure 4.

Denoting the proportion female in each study as p_i, the proportion of women in the overall population which we have targeted by use of weights ω_i is con-

strained to be

$$p_\omega = \frac{\sum_i p_i \eta_i \omega_i}{\sum_i \eta_i \omega_i},$$

and the corresponding meta-analytic estimator is

$$\frac{\sum_i \hat{\beta}_i \omega_i \eta_i \phi_i}{\sum_i \omega_i \eta_i \phi_i},$$

where $\eta_i \phi_i \propto \sigma_i^{-2}$. Of course, there are many such weights $\boldsymbol{\omega}$ that will achieve the constraint on p_ω. Using Lagrange multipliers (details omitted) it can be shown that the choice which minimizes the variance of the meta-analytic estimator sets

$$\omega_i \propto (p_i - p_\omega)\phi_i^{-1} - \sum_{i'} \frac{(p_{i'} - p_\omega)\eta_{i'}}{\sum_j (p_j - p_\omega)\eta_j}(p_{i'} - p_\omega)\phi_i^{-1}.$$

We see that weights are higher for studies for which quantity $(p_i - p_w)\phi^{-1}$ is far from the average such quantity, but where this average is weighted by the relative size of $(p_i - p_w)\eta_i$. In this way the weights accomodate both the per-observtion information, and the relative size of the studies, whilst satisfying the constraint on p_ω.

The approach is illustrated in Figure 4, using the data from Figure 1. When p_ω corresponds to the proportion in the unweighted amalgamation of all the studies, the estimate is identical to the standard fixed-effects approach. Elsewhere the estimate changes little, relative to its statistical uncertainty, suggesting that the effect size differs little as we consider populations with different sex balances.

Another example of reweighting based on elicited prior opinion regarding the relevance of each study to a particular scenario is given by Turner et al. (2009).

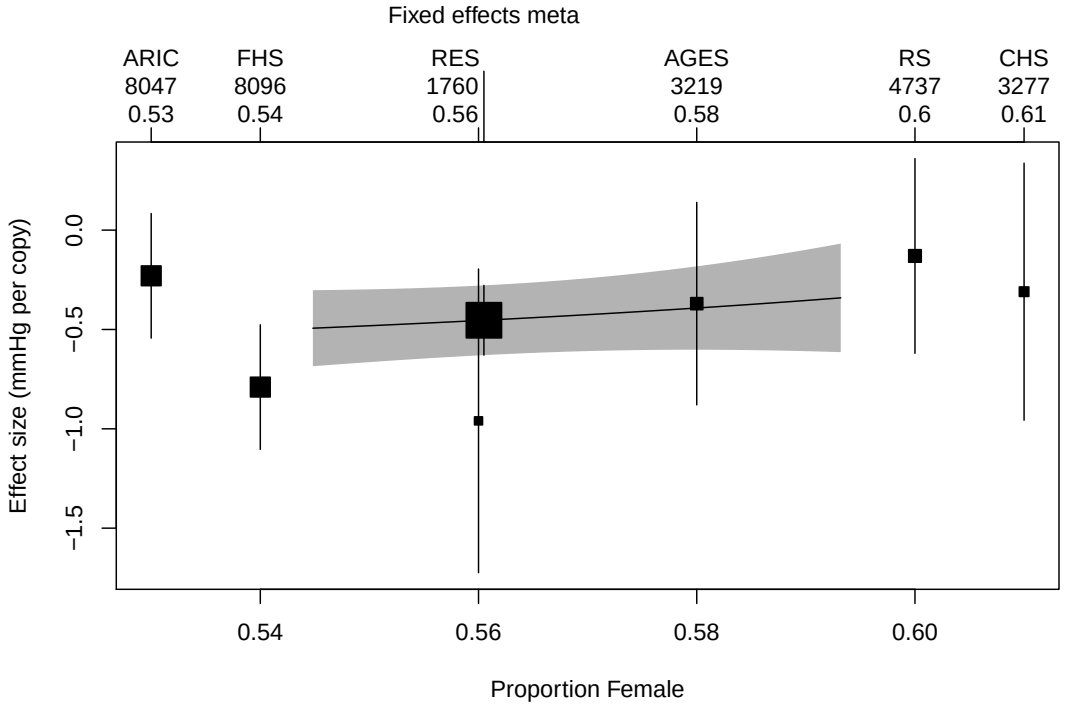


Figure 4: Estimates from overall populations chosen to have specific proportion female, using data from Levy et al. (2009). The individual study effects from Figure 1 are plotted against their corresponding proportion female; the sample sizes and proportions female are given above the plot. For a selected range of proportions female, the point estimate and 95% confidence interval are shown, for the parameter that is estimated by weighted linear regression of blood pressure on genotype, adjusting for study.