Published in final edited form as:
Stat Biopharm Res. 2010 August 1; 2(3): 300–309. doi:10.1198/sbr.2009.0067.

Measuring Study-Specific Heterogeneity in Meta-Analysis: Application to an Antecedent Biomarker Study of Alzheimer's Disease

Chengjie Xiong,
TKKK, Division of Biostatistics, Washington University, St. Louis, MO 63110

J. Philip Miller, and
Professor of Biostatistics, Washington University, St. Louis, MO 63110

John C. Morris
Friedman Distinguished Professor of Neurology, Washington University, St. Louis, MO 63110

Abstract

This article proposes several new indices that measure the heterogeneity for individual studies in a meta-analysis. These indices directly assess how inconsistent an individual study is compared to the rest of studies used in the meta-analysis, that is, how much impact the specific study has on the scientific conclusion of the meta-analysis and further on the generalization of the conclusion. The proposed indices can be intuitively interpreted as the proportion of total variance from all studies in a meta-analysis that can be accounted for by the heterogeneity from specific studies. Further, each proposed index over all the studies sums to the collective measure of heterogeneity for the meta-analysis. Therefore our proposed study-specific indices of heterogeneity can be regarded as a generalization of the collective index of heterogeneity in meta-analyses proposed by various authors. We examine the difference among the proposed study-specific measures of heterogeneity and assess the variation associated with each proposed index of heterogeneity through a large simulation study. Finally, we demonstrate the proposed methodology by assessing the effect of individual studies on the overall estimate to the difference of an antecedent biomarker of Alzheimer's disease (AD) between different Apolipoprotein E (ApoE) genotypes.

Keywords

Random effects models

1. Introduction

Making decisions on medicine and health care policies is a very complicated process in which existing scientific evidence plays a crucial role. Medical practitioners and their patients make decisions within the context of a rapidly changing body of scientific evidence on medicine and a health care system that influences the availability, accessibility, and cost of diagnostic tests and therapies (Sackett and Haynes 1995). Timely, useful evidence from the biomedical literature should be an integral component of clinical and medical decision making. The importance of basing medical practice more firmly on the results of existing scientific evidence through systematic reviews was starkly demonstrated by a paper in the

Correspondence: chengjie@wustl.edu.

early 1990s, which compared the results of meta-analyses of treatment trials for people who have suffered a heart attack with the recommendations of experts published in review articles and textbooks over the same time period. This showed a significant divergence between the recommendations and the meta-analytic summaries of the trials. Ineffective treatments were being recommended, and highly effective treatments were not. There were also significant time delays between the publication of the studies and changes in the recommendations of the experts (Antman et al. 1992). As a result, lives that could have been saved were lost, and resources were wasted.

Systematic reviews are very useful medical decision-making tools because they objectively summarize large amounts of information, identify gaps in medical research and evidence, and identify beneficial or harmful interventions. Clinicians can use systematic reviews to guide their patient care. Consumers and patients as well as policymakers can use systematic reviews to help make health care decisions. Systematic reviews provide convincing and reliable evidence relevant to many aspects of medical and biological research and health care (Egger and Smith 1997), especially when the results of individual studies they include show clinically important effects of comparable magnitude. Such reviews aim to comprehensively identify and assess all studies relevant to a given scientific question, and meta-analysis has been the major statistical methodology for the quantitative synthesis of study results. Many methods for meta-analysis are available, and the most popularly applied in the medical research focus on the optimum combination of published summary statistics in some form of weighted averages (DerSimonian and Laird 1986; Egger, Smith, and Phillips 1997; Whitehead and Whitehead 1991). Usually, each study is given a weight according to the precision of its results on summary statistics. Studies with good precision are weighted more heavily than studies with greater uncertainty. The variance for the overall estimate of the parameter under study in meta-analyses is in general from two different sources, one is associated with the individual studies (i.e., the within-study variance), and the other is associated with the possible difference between different studies (i.e., between-study variance). When the between-study variance is assumed to be 0, each study is simply weighted according to its own variance. This approach characterizes a fixed effects model which is exemplified by the Mantel-Haenszel method (Mantel and Haenszel 1959; Laird and Mosteller 1990) or the Peto method (Yusuf 1985). When the between-study variance is not zero, methods which incorporate a between-study component of variation for the overall effect under estimation are based on random effects models (Laird and Mosteller 1990). The between-study variance represents the excessive variation in observed individual study effects over that expected from the imprecision of results within each study. Heterogeneity in a meta-analysis refers to the *between*-study variance of each individual study when the overall mean of the random effects is estimated. Fixed effects and random effects model for general continuous outcome and specific survival outcomes have been described by Hedges and Olkin (1985); Earle and Wells (2000); Srinivasan and Zhou (1993); and Parmar, Torri, and Stewart (1998).

When individual studies used in a meta-analysis have very differing results, however, the results from systematic reviews may be less convincing and reliable. In an attempt to establish whether study results are consistent, reports on a meta-analysis commonly present a statistical test of heterogeneity among studies used in the meta-analysis. This test seeks to determine whether there are genuine differences underlying the results of the studies, or whether the variation in these results is compatible with chance alone (i.e., homogeneity). A common statistical test used for this purpose is the Cochran's chi-squared test or the *Q*-test (Whitehead and White-head 1991; Cochran 1954). It has been widely realized, however, that this test has poor power when the number of studies in a meta-analysis is small, and excessive power to detect clinically insignificant heterogeneity when there are too many studies (Higgins and Thompson 2002; Hardy and Thompson 1998).

Addressing statistical heterogeneity of studies is one of the most fundamental aspects of many systematic reviews. The interpretative aspects of statistical inferences from a meta-analysis depend on the degree of heterogeneity of the studies used in the meta-analysis. Because the heterogeneity may determine the extent to which the conclusions of a meta-analysis can be generalized, it is important to quantify the extent of heterogeneity among a collection of studies. Realizing the potential limitations of statistical tests to characterize the degree of heterogeneity in a meta-analysis, Higgins and Thompson (2002) proposed new measures of the extent of heterogeneity in a meta-analysis that overcome the shortcomings of existing measures. Their focus is on the impact of heterogeneity on the results of a meta-analysis and therefore, on the degree to which scientific conclusions might be generalized to situations outside those investigated in the studies at hand. Their measures are easily interpreted by nonstatisticians as the proportion of variation that was explained by the difference among studies. Further, these measures do not intrinsically depend on the number of studies or the type of outcome data, therefore offering the possibility that statistical heterogeneity can be compared across different meta-analyses with differing numbers of studies and types of outcome data. Because of the fact that their proposed measures of heterogeneity in a meta-analysis measure the overall or collective heterogeneity within the group of studies used in a meta-analysis, the interpretation of the index on heterogeneity has to refer to the collection of studies used in the meta-analysis.

Often times, however, a scientifically important question to be answered in a meta-analysis is how inconsistent one specific study is compared to the rest of studies used in the meta-analysis, that is, how much impact each individual study has on the scientific conclusion of the meta-analysis and further on the generalization of the conclusion. Because heterogeneity comes about due to the fact that the effects under study in the population which the studies represent are not the same, it is important to understand the sources and possible explanations of the heterogeneity, including study sample characteristics, the design and analytic features used to report results, and the scientific interpretations of the study results. All these can only be facilitated when heterogeneity of individual studies can be directly measured in comparison to the rest of the studies in the meta-analyses.

In this article, we propose several new indices that measure the specific inconsistency for an individual study as compared to the rest of studies used in a meta-analysis. We seek to develop indices that will measure the study-specific degree of inconsistency in such a way that sheds light on the degree of contribution of this specific study to the overall conclusion of the meta-analysis. The proposed methodology can be regarded as a generalization of the collective index of heterogeneity proposed by Higgins and Thompson (2002). We also examine the difference among the proposed study-specific measures of heterogeneity and study the variation of each proposed measure when a large number of simulated meta-analyses are conducted. Finally, we demonstrate our proposed methodology by presenting an example to study possible biomarkers that can be used to identify subjects with high risk of developing Alzheimer's disease (AD) when they are still cognitively normal.

2. Indices of Study-Specific Heterogeneity in a Meta-Analysis

We assume that a total of k studies are used in a meta-analysis to address a scientific question as represented by parameter θ. Let $\hat{\theta}_i$ be the estimate from the i th study and $\widehat{\sigma}_i^2$ be the associated estimate to the variance. Let $w_i=1/\widehat{\sigma}_i^2$ denote the precision of the estimate. In a classic fixed effect meta-analysis, θ_i's are assumed identical and a summary estimate, $\hat{\theta}$, is computed to the common parameter as a weighted average of the study specific estimates, using the precisions as weights:

$$\widehat{\theta}=\frac{\sum_{i=1}^{k} w_i\widehat{\theta_i}}{\sum_{i=1}^{k} w_i}.$$

The variance of the summary estimate is given by

$$\widehat{\sigma}^2_{\widehat{\theta}}=\frac{1}{\sum_{i=1}^{k} w_i}.$$

A random effects meta-analysis can be conceptualized by incorporating a random effect to account for the between-study variation, $N(0, \tau^2)$, into the estimated study-specific parameters, in addition to the within-study random variation, $N(0, \sigma_i^2)$. The summary estimate to the mean parameter across the distribution of the studies, $\hat{\theta}_r$, has exactly the same form as above, but with weights replaced by

$$\widehat{w_i^*}=\frac{1}{(w_i^{-1}+\tau^2)}.$$

The estimated variance of the summary estimate is now given by

$$\widehat{\sigma}^2_{\widehat{\theta}_r}=\frac{1}{\sum_{i=1}^{k}\widehat{w_i^*}}.$$

A test of homogeneity of the θ_i is given by

$$Q=\sum_{i=1}^{k} w_i(\widehat{\theta_i}-\widehat{\theta})^2,$$

which has a chi-squared distribution with $k-1$ degrees of freedom under the assumption of homogeneity in the fixed effects model. A method of moment estimate to τ^2 can be obtained as

$$\widehat{\tau}^2=\frac{Q-(k-1)}{\sum_{i=1}^{k} w_i-\frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}}. \quad (1)$$

Notice that in the fixed effect model, the assumptions of known sampling variances and normally distributed effect size estimates are usually approximations based on the large sample theory of maximum likelihood estimates. Further, the random-effects model weights ignore the uncertainty in $\hat{\tau}^2$. The meta-analyses results are therefore only valid with large within-study sample sizes to approximate known sampling variances and normally

distributed estimates and large number of studies (i.e., k) to reduce the imprecision in the estimate of τ^2.

Higgins and Thompson (2002) proposed a simple index to quantify the overall heterogeneity among studies in a meta-analysis:

$$I^2 = \frac{\tau^2}{\tau^2 + \sigma^2},$$

where σ^2 is the shared within-study variance among individual studies, or when the studies have differing within-study variations, the "typical" within-study variance in the terms of Higgins and Thompson (2002). This intuitive definition of the heterogeneity has several major advantages as compared to the standard statistical test based on Q. First, the measure does not inherently depend on the number of studies in the meta-analysis. Second, the measure is not specific to a particular metric of treatment effect and therefore can be applied similarly irrespective of the type of outcome data (e.g., dichotomous, continuous, and survival). Third, the measure is easy to compute and has a very appealing interpretation as the percentage of the total variation across studies due to heterogeneity.

The estimation of overall heterogeneity among studies in a meta-analysis requires the estimate to both the between-study variation and the "typical" within-study variance. Higgins and Thompson (2002) used the following estimator

$$\widehat{\sigma}^2_{HT} = \frac{(k-1)\sum_{i=1}^{k} w_i}{\left(\sum_{i=1}^{k} w_i\right)^2 - \sum_{i=1}^{k} w_i^2}$$

to estimate the "typical" within-study variance, and derived the index of overall heterogeneity

$$I^2_{HT} = \frac{\widehat{\tau}^2}{\widehat{\tau}^2 + \widehat{\sigma}^2_{HT}} = \frac{Q-(k-1)}{Q}. \quad (2)$$

Takkouche, CadarsoSurez, and Spiegelman (1999) suggested another estimate to the "typical" within-study variance σ^2 by taking the reciprocal of the arithmetic mean weights:

$$\widehat{\sigma}^2_T = \frac{k}{\sum_{j=1}^{k} w_j}.$$

This gives another index of overall heterogeneity

$$I^2_T = \frac{\widehat{\tau}^2}{\widehat{\tau}^2 + \widehat{\sigma}^2_T} = \frac{Q-(k-1)}{Q+1-\frac{k\sum_{j=1}^{k} w_j^2}{\left(\sum_{j=1}^{k} w_j\right)^2}}. \quad (3)$$

Taking the simple arithmetic average of the within-study variances

$$\widehat{\sigma}_s^2 = \frac{\sum_{j=1}^{k} \frac{1}{w_j}}{k}$$

to estimate the "typical" within-study variance results in one more index of overall heterogeneity

$$I_s^2 = \frac{\widehat{\tau}^2}{\widehat{\tau}^2 + \widehat{\sigma}_s^2} = \frac{Q-(k-1)}{Q-k+1+\sum_{j=1}^{k} \frac{1}{w_j} \frac{\left[\left(\sum_{j=1}^{k} w_j\right)^2 - \sum_{j=1}^{k} w_j^2\right]}{k\sum_{j=1}^{k} w_j}}. \quad (4)$$

We follow the convention that in all these proposed indices of heterogeneity, they are set to 0 if $Q \le (k-1)$.

For a specific study i, we accordingly propose three different indices to measure its heterogeneity from the collection of studies used in the meta-analysis:

$$I_{HT}^2(i) = \frac{w_i(\widehat{\theta}_i - \widehat{\theta})^2 - \delta_i}{\sum_{j=1}^{k} w_j(\widehat{\theta}_j - \widehat{\theta})^2}, \quad (5)$$

$$I_T^2(i) = \frac{w_i(\widehat{\theta}_i - \widehat{\theta})^2 - \delta_i}{Q+1-\frac{k\sum_{j=1}^{k} w_j^2}{\left(\sum_{j=1}^{k} w_j\right)^2}}, \quad (6)$$

and

$$I_s^2(i) = \frac{w_i(\widehat{\theta}_i - \widehat{\theta})^2 - \delta_i}{Q-k+1+\sum_{j=1}^{k} \frac{1}{w_j} \frac{\left[\left(\sum_{j=1}^{k} w_j\right)^2 - \sum_{j=1}^{k} w_j^2\right]}{k\sum_{j=1}^{k} w_j}}, \quad (7)$$

where

$$\delta_i = 1 - \frac{w_i}{\left(\sum_{j=1}^{k} w_i\right)}.$$

If all within-study variations are exactly the same, then $\delta_i = (k-1)/k$ and $I_{HT}^2(i) = I_T^2(i) = I_s^2(i)$. We also follow the convention that if the numerator is negative in these indices, that is, $\omega_i(\hat{\theta}_i - \hat{\theta})^2 \le \delta_i$, then $I_{HT}^2(i) = I_T^2(i) = I_s^2(i) = 0$.

3. Properties of the Proposed Study-Specific Measures of Heterogeneity

From Equation (1) of Higgins and Thompson (2002), the expected value of Q statistic is

$$E(Q)=\tau^2\left[\sum_{i=1}^{k}w_i-\frac{\sum_{i=1}^{k}w_i^2}{\sum_{i=1}^{k}w_i}\right]+k-1.$$

If there is no heterogeneity among studies, that is, $\tau^2 = 0$, then $E(Q) = k - 1$. A similar mathematical derivation gives

$$E\left(w_i(\widehat{\theta}_i-\widehat{\theta})^2\right)=\frac{w_i\tau^2}{\left(\sum_{j=1}^{k}w_j\right)^2}\left[\left(\sum_{j=1}^{k}w_j-w_i\right)^2+\left(\sum_{j=1}^{k}w_j^2-w_i^2\right)\right]+1-\frac{w_i}{\left(\sum_{j=1}^{k}w_j\right)}.$$

If there is no heterogeneity among studies, that is, $\tau^2 = 0$, then

$$E\left(w_i(\widehat{\theta}_i-\widehat{\theta})^2\right)=\delta_i.$$

This also results in another method of moment estimate of τ^2 as

$$\widehat{\tau}_i^2=\frac{w_i(\widehat{\theta}_i-\widehat{\theta})^2-\delta_i}{\frac{w_i}{\left(\sum_{j=1}^{k}w_j\right)^2}\left[\left(\sum_{j=1}^{k}w_j-w_i\right)^2+\sum_{j=1}^{k}w_j^2-w_i^2\right]}.$$

Notice that the denominator of all the proposed overall and study-specific measures of heterogeneity is the unconditional variance of the estimated effect from a typical study in the meta-analysis, which contains additive components due to the within-study variance (i.e., from between-patient variation within the study) and the between-study variation (i.e., heterogeneity).

By Schwartz's inequality (Noble and Daniel 1977),

$$\left(\sum_{j=1}^{k}w_j\right)^2\le k\sum_{j=1}^{k}w_j^2,$$

and

$$k^2=\left(\sum_{j=1}^{k}\sqrt{w_j}\frac{1}{\sqrt{w_j}}\right)^2 \le \sum_{j=1}^{k}w_j\sum_{j=1}^{k}\frac{1}{w_j}.$$

It then follows that

$$I_S^2 \le I_T^2$$

and

$$I_{HT}^2 \le I_T^2.$$

Similarly, it is clear that for any study i,

$$I_S^2(i) \le I_T^2(i)$$

and

$$I_{HT}^2(i) \le I_T^2(i).$$

Notice that if $I^2(i) > 0$ for all i, then

$$I_{HT}^2=\sum_{i=1}^{k}I_{HT}^2(i),$$
$$I_T^2=\sum_{i=1}^{k}I_T^2(i),$$

and

$$I_S^2=\sum_{i=1}^{k}I_S^2(i),$$

that is, the total heterogeneity in a meta-analysis can be partitioned as the simple sum of these from individual studies. Therefore, the intuitive interpretation of overall heterogeneity I^2 can be inherited to interpret the study-specific measures of heterogeneity $I^2(i)$ as the proportion of total variance that can be accounted for by the heterogeneity from study i.

Notice also that the study specific measures $I_S^2(i)$, $I_T^2(i)$, and $I_{HT}^2(i)$ depend on k and decrease when k increases. Therefore, as k increases, less and less of the total heterogeneity can be attributed to any single study.

4. Distributions and Comparisons of the Study-Specific Indices of Heterogeneity

Although mathematically, $I_S^2(i) \le I_T^2(i), I_{HT}^2(i) \le I_T^2(i)$ for each study i, it is impor tant to understand how different these measures are when they are used to measure the study-specific heterogeneity in a meta-analysis and how much variation each index has when a large number of meta-analyses are conducted. Given the fact that when all studies have exactly the same degree of within-study variation, that is, when all ω_i's are the same, these measures are identical to each other, we anticipate that these measures will be close to each other when the difference among within-study variations is relatively small.

We performed a simulation study to look at the performance of our proposed indices of study-specific heterogeneity. For this purpose, we first generated a specific study whose parameter estimate is generated by the random effect model with the between-study component following the normal distribution $N(5, \tau^2)$ with τ^2=0, 1, 4, through a linear transformation of the SAS function RANNOR (SAS 1999). The within-study precision for the specific study is among one of the three possible values: 0.5 + υ, or 0.5 + 2υ, or 0.5 + 3υ for a range of υ = 0, 0.5, and 2.0. The proposed study-specific measures of heterogeneity are computed for this specific study in each simulated meta-analysis. In addition to this specific study, other 3s (for s = 4 and 8) studies in the meta-analysis are generated by the same random effect model but with within-study precisions equally distributed among the three possible values: 0.5 + υ, or 0.5 + 2υ, or 0.5 + 3υ for a range of υ, that is, s studies have one of the three possible within-study precision values. Therefore, the total number of studies used in each meta-analysis is k = 3s + 1 where s was chosen as 4 and 8. For each possible value of τ^2, s, and v, 1000 independent simulated meta-analyses were performed such that study estimates for the specific study and the other 3s studies were independently generated across 1000 meta-analyses. Table 1 presents the mean and standard error for the three proposed measures of study-specific heterogeneity over 1000 simulated meta-analyses as a function of τ^2, k, and v (notice that v is a measure of heterogeneity among the study precisions). In addition, Table 1 also presents the true overall measures of heterogeneity for each scenario.

Notice that our simulation results in Table 1 cover a wide range of true degree of heterogeneity with the true index from 0% to almost 95%. From our simulated meta-analyses, it is clear that three different measures of overall and study specific heterogeneity are very consistent within the specified ranges of parameters. In fact, under the assumption that the three measures of heterogeneity are estimating the same underlying heterogeneity, we computed the intraclass correlation coefficient (ICC) (Shrout and Fleiss 1979) over 1000 simulated meta-analyses for each choice of τ^2, k, and v. All these computed ICCs were at least 0.99, indicating extremely high consistency among these measures. When $\tau^2 = 0$, there is no heterogeneity across studies in the meta-analyses, that is, $I_{HT}^2 = I_T^2 = I_S^2 = 0$, which should then imply that $I_{HT}^2(i) = I_T^2(i) = I_S^2(i) = 0$ for each individual study i. However, because of a positive probability that $\omega_i (\hat{\theta}_i - \hat{\theta})^2 \le \delta_i$, we made the convention to define $I_{HT}^2(i) = I_T^2(i) = I_S^2(i) = 0$ in this case. This truncation therefore leads to a possible positive bias. The results in Table 1 when $\tau^2 = 0$ present the estimates to the degree of the positive bias due to the truncation to 0.

5. Application to an Antecedent Biomarker Study of Alzheimer's Disease

Alzheimer's disease (AD) is a highly complex and multi-factorial progressive neurological disease that results in the irreversible loss of neurons in one or multiple regions of the brain.

We present an application to our proposed overall and study-specific measures of heterogeneity to study possible biomarkers that can be used to identify individuals with high risk of developing Alzheimer's disease (AD) when they are still cognitively normal. Recent research advances in Alzheimer's disease have found Apolipoprotein E4 (ApoE4) alleles as a genetic risk factor of AD (Myers 1996). Although the pathological hallmarks of AD are the neurofibrillary tangles and the senile plaques in the brain (Braak and Braak 1991; McKeel et al. 2004; Fagan et al. 2007), the diagnosis of AD in living patients is still largely a clinical judgment based on careful neurological and/or neuropsychological examination combined with results from other clinical tests. Therefore, the search for biomarkers that can be used to differentiate AD from normal aging remains one of the primary research activities in AD. In several publications (Fagan et al. 2007; Sunderland et al. 2003), individuals with AD were found to have decreased level of cerebrospinal fluid (CSF) β-amyloid$_{42}$ as compared to individuals with normal aging. Because AD is a progressive neurodegenerative disorder that leads to the irreversible death of brain cells, it is important to assess the potential of the CSF biomarker to identify individuals that are at high risk of AD while they are still cognitively normal. The importance of such antecedent biomarkers is further highlighted by the fact that no pharmaceutical treatments are effective for the disease's later stages. We chose to study whether CSF β-amyloid$_{42}$ is decreased among individuals of normal aging who are ApoE4 positive as compared to these who are ApoE4 negative. Although many publications have compared CSF β-amyloid$_{42}$ level between individuals with AD and these with normal aging (Fagan et al. 2007; Sunderland et al. 2003), very few have actually reported CSF β-amyloid$_{42}$ as a function of ApoE4 status among subjects who were still cognitively normal. As a matter of fact, our comprehensive MEDLINE search identified a total of only six published studies on CSF β-amyloid$_{42}$ during the period of 1990 to 2007 which actually reported summary statistics as a function of ApoE4 status for individuals who were not demented (Sunderland 2004; Jensen et al. 1999; Andreasen et al. 1999; Tapiola et al. 2000; Riemenschneider et al. 2000; Prince et al. 2004). The summary statistics reported from these six published studies are presented in Table 2 [summary statistics from the study by Prince et al. (2004) was obtained through eye-balling because only a graphical presentation on summary statistics was available in the publication].

Based on our proposed methodology and a random effect model, the pooled estimate to the mean difference of CSF β-amyloid$_{42}$ between individuals of normal aging who are ApoE4 positive and those who are ApoE4 negative is −31.69 pg/mL, and an asymptotic 95% confidence interval estimate to the mean difference of CSF β-amyloid$_{42}$ is from −128.93 pg/mL to 65.56 pg/mL, suggesting a nonsignificant difference at a 5% significance level. The measures of overall heterogeneity from this meta-analysis are estimated as $I^2_{HT}=0.56$, $I^2_T=0.66$, and $I^2_S=0.20$, respectively, indicating from low to moderate degree of heterogeneity among studies used in the meta-analysis (Higgins et al. 2003). If the heterogeneity is ignored in the meta-analysis, that is, the between-study variance τ^2 is assumed as 0 (therefore $I^2_{HT}=I^2_T=I^2_S=0$), then a fixed effect model would be used for the meta-analysis. The estimated overall mean difference of CSF β-amyloid$_{42}$ between individuals of normal aging who are ApoE4 positive and those who are ApoE4 negative under the fixed effect model is −45.35 pg/mL. An asymptotic 95% confidence interval estimate to the mean difference of CSF β-amyloid$_{42}$ under the fixed effect model is from −74.89 pg/mL to −15.82 pg/mL, suggesting a statistically significant difference at a 5% significance level on CSF β-amyloid$_{42}$ between individuals of normal aging who are ApoE4 positive and those who are ApoE4 negative. This discrepancy on the statistical inference between the fixed effect model and the random effect model is *partly* due to the fact that one approach (i.e., the random effect model) takes into account of heterogeneity across studies whereas the other (i.e., the fixed effect model) ignores such heterogeneity, suggesting the importance to measure the heterogeneity in meta-analyses when it does exist. The fixed-effects model

provides a conditional inference about the set of studies included in the meta-analysis, while the random-effects model provides an unconditional inference about a hypothetical population of studies (from which the included studies are assumed to be a random sample). Either model provides the appropriate inferences under the specific assumptions under the model (Hedges and Vevea 1998).

Columns 3 to 5 of Table 3 display the study-specific measures of heterogeneity for all six studies. All three indices indicated that the study by Prince et al. (2004) has the largest heterogeneity from the rest of studies. In fact, the study by Prince et al. (2004) alone accounts for from 12% to 40% of overall heterogeneity in the meta-analysis. The last column of Table 3 presents the pooled estimate to the mean difference of CSF β-amyloid$_{42}$ between individuals of normal aging who are ApoE4 positive and those who are ApoE4 negative when one study is excluded from the meta-analysis with a random effect model. When the study by Prince et al. (2004) was excluded from the meta-analysis, the pooled estimate to the mean difference of CSF β-amyloid$_{42}$ was −9.13 pg/mL, giving the largest deviation from the pooled estimate when all six studies were included in the meta-analysis.

6. Discussion

We proposed several new indices that measure the heterogeneity for individual studies as compared to the rest of studies used in a meta-analysis. By estimating the “typical” within-study precisions, we developed these indices that measure the degree of inconsistency among studies by their impact to the overall conclusion of the meta-analysis. The proposed methodology can be regarded as a generalization of the collective index of heterogeneity proposed by Higgins and Thompson (2002). We assessed the variation associated with each proposed index of heterogeneity through a large simulation study. We also examined the difference among the proposed study-specific measures of heterogeneity and found that these indices provided quite consistent results in measuring the study-specific heterogeneity in the simulated meta-analyses. Finally, we demonstrated our proposed methodology by presenting a real world application to study a CSF biomarker that can be used to identify individuals with high risk of developing Alzheimer’s disease (AD) when they are still cognitively normal. We further identified these studies that have the most heterogeneity in this example, and assessed their individual effect to the overall estimate on the effect size of ApoE4 genotypes.

Our proposed study-specific measures of heterogeneity directly assess how inconsistent one specific study is compared to the rest of studies used in the meta-analysis, that is, how much impact the specific study has on the scientific conclusion of the meta-analysis and further on the generalization of the conclusion. Further each proposed index has another simple appealing property that its sum over all the studies used in the meta-analysis is the same as the overall measure of heterogeneity for the meta-analysis. This simple property allows the interpretation of study-specific measures of heterogeneity within the context of overall measures of heterogeneity in a meta-analysis and therefore inherits the appealing conceptualization that the study-specific measures represent the proportion of total variance across studies that can be accounted for by the heterogeneity from specific studies.

Addressing statistical heterogeneity of studies is one of the most important aspects of many systematic reviews. The interpretative aspects of statistical inferences from a meta-analysis depend on the degree of heterogeneity of the studies used in the meta-analysis. Because heterogeneity comes about due to the fact that the effects under study in the populations which the studies represent are not the same, it is important to understand the sources and possible explanations of the heterogeneity. When individual studies used in a meta-analysis have very differing results, knowing the exact contribution of individual studies to the total

heterogeneity becomes the first step to understand the sources of heterogeneity. This information can not only identify studies with the largest heterogeneity in a meta-analysis but also help more careful assessments on the individual studies to make sure they are consistent in patient characteristics and study designs as well as analytic approaches. If there is enough evidence suggesting that the heterogeneity of a specific study is extremely large compared to other studies and mainly due to different patient populations or different study designs or less-than-optimal analytic approaches, the protocol of the meta-analysis may be revised to exclude the study, or meta-analytic results with and without the study may be both reported to allow an assessment on the impact of the single study on the scientific conclusions. In fact, with our proposed study-specific indices of heterogeneity, it becomes possible that future meta-analyses report the study-specific heterogeneity indices to give an estimate to the proportion of total variance in the reported effect sizes that can be accounted for by individual studies.

Acknowledgments

Dr. Xiong's work was supported by grant K25 AG025189 from the National Institute on Aging. Financial support for this study was also provided in part by National Institute on Aging grants AG003991, AG005681, and AG026276 for Chengjie Xiong, J. Philip Miller, and John C. Morris.

References

Andreasen N, Hesse C, Davidsson P, et al. Cerebrospinal Fluid β-amyloid$_{(1–42)}$ in Alzheimer's Disease: Differences between Early- and Late-onset Alzheimer's Disease and Stability during the Course of Disease. Archives of Neurology 1999;56:673–680. [PubMed: 10369305]

Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A Comparison of Results of Meta-analyses of Randomized Control Trials and Recommendations of Clinical Experts: Treatment for Myocardial Infarction. Journal of the American Medical Association 1992;268:240–248. [PubMed: 1535110]

Braak H, Braak E. Neuropathologic Staging of Alzheimer-related Changes. Acta Neuropathologica 1991;82:239–259. [PubMed: 1759558]

Cochran WG. The Combination of Estimates from Different Experiments. Biometrics 1954;10:101–129.

DerSimonian R, Laird NM. Meta-analysis in Clinical Trials. Controlled Clinical Trials 1986;7:177–188. [PubMed: 3802833]

Earle CC, Wells GA. An Assessment of Methods to Combine Published Survival Curves. Medical Decision Making 2000;20:104–111. [PubMed: 10638543]

Egger M, Smith GD. Meta-analysis: Potentials and Promise. British Medical Journal 1997;315:1371–1374. [PubMed: 9432250]

Egger M, Smith GD, Phillips AN. Meta-analysis; Principles and Procedures. British Medical Journal 1997;315:1533–1537. [PubMed: 9432252]

Fagan AM, Roe CM, Xiong C, Mintun MA, Morris JC, Holtzman DM. Cerebrospinal Fluid tau/β-Amyloid$_{42}$ Ratio as a Prediction of Cognitive Decline in Nondemented Older Adults. Archives of Neurology 2007;64:343–349. [PubMed: 17210801]

Hardy RJ, Thompson SG. Detecting and Describing Heterogeneity in Meta-analysis. Statistics in Medicine 1998;17:841–856. [PubMed: 9595615]

Hedges, LV.; Olkin, I. Statistical Methods for Meta-Analysis. Orlando, FL: Academic Press; 1985.

Hedges LV, Vevea JL. Fixed- and Random-effects Models in Meta-analysis. Psychological Methods 1998;3:486–504.

Higgins JP, Thompson SG. Quantifying Heterogeneity in a Meta-analysis. Statistics in Medicine 2002;21:1539–1558. 2, 3, 4, 8. [PubMed: 12111919]

Higgins JPT, Thompson SG, Deeks JJ, et al. Measuring Inconsistency in Meta-analyses. British Medical Journal 2003;327:557–560. [PubMed: 12958120]

Jensen M, Schroder J, Blomberg M, et al. Cerebrospinal Fluid *Aβ*42 is Increased Early in Sporadic Alzheimer's Disease and Declines with Disease Progression. Annals of Neurology 1999;45:504–511. [PubMed: 10211475]

Laird NM, Mosteller F. Some Statistical Methods for Combining Experimental Results. International Journal of Technology Assessment in Health Care 1990;6:5–30. [PubMed: 2361819]

Mantel N, Haenszel W. Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease. Journal of the National Cancer Institute 1959;22:719–748. [PubMed: 13655060]

McKeel DW Jr, Price JL, Miller JP, Grant EA, Xiong C, Berg L, Morris JC. Neuropathologic Criteria for Diagnosing Alzheimer Disease in Persons with Pure Dementia of Alzheimer Type. Journal of Neuropathology and Experimental Neuropathology 2004;63(10):1028–1037.

Myers RH, Schaefer EJ, Wilson PWF, et al. Apolipoprotein E *ε*4 Association with Dementia in a Population-based Study: The Framingham Study. Neurology 1996;46:673–677. [PubMed: 8618665]

Noble, B.; Daniel, JW. Applied Linear Algebra. Englewood Cliffs, NJ: Prentice-Hall Inc; 1977.

Parmar MKB, Torri V, Stewart LA. Extracting Summary Statistics to Perform Meta-analyses of Published Literature for Survival Endpoints. Statistics in Medicine 1998;17:2815–2834. [PubMed: 9921604]

Prince JA, Zetterberg H, Andreasen N, et al. APOE *ε* 4 allele Is Associated with Reduced Cerebrospinal Fluid Levels of A*β*42. Neurology 2004;62:2116–2118. 7, 8. [PubMed: 15184629]

Riemenschneider M, Schmolke M, Lautenschlager N, et al. Cerebrospinal beta-amyloid$_{(1–42)}$ in Early Alzheimer's Disease: Association with Apolipoprotein E Genotype and Cognitive Decline. Neuroscience Letters 2000;284:85–88. [PubMed: 10771168]

Sackett DL, Haynes RB. On the Need for Evidence-based Medicine. Evidence-Based Medicine 1995;1:5–6.

SAS Institute, Inc. SAS Language (Version 8). Cary, NC: 1999.

Shrout PE, Fleiss JL. Intraclass Correlations: Uses in Assessing Rater Reliability. Psychological Bulletin 1979;86:420–428. [PubMed: 18839484]

Srinivasan C, Zhou M. A Note on Pooling Kaplan–Meier Estimators. Biometrics 1993;49:861–864.

Sunderland T, Linker G, Mirza N, et al. Decreased *β*-amyloid$_{1–42}$and Increased tau Levels in Cerebrospinal Fluid of Patients with Alzheimer's Disease. Journal of the American Medical Association 2003;289:2094–2103. [PubMed: 12709467]

Sunderland T, Mirza N, Putnam KT, et al. Cerebrospinal Fluid *β*-amyloid$_{1–42}$and tau in Control Subjects at Risk for Alzheimer's Disease: The Effect of ApoE *ε*4 Allele. Biological Psychiatry 2004;56:670–676. [PubMed: 15522251]

Takkouche B, CadarsoSurez C, Spiegelman D. Evaluation of Old and New Tests of Heterogeneity in Epidemiologic Meta-analysis. American Journal of Epidemiology 1999;150:206–215. [PubMed: 10412966]

Tapiola T, Pirttila T, Mehta PD, et al. Relationship between ApoE Genotype and CSF *β*-amyloid$_{(1–42)}$ and tau in Patients with Probable and Definite Alzheimer's Disease. Neurobiology of Aging 2000;21:735–740. [PubMed: 11016543]

Whitehead A, Whitehead J. A General Parametric Approach to the Meta-analysis of Clinical Trials. Statistics in Medicine 1991;10:1665–1677. [PubMed: 1792461]

Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta Blockade during and after Myocardial Infarction: An Overview of the Randomized Trials. Progress in Cardiovascular Diseases 1985;27:335–371. [PubMed: 2858114]

Table 1

Mean (standard error (SE)) of study-specific heterogeneity (in %) from 1000 simulated meta-analyses (k = the number of studies in meta-analyses, τ^2 = between-study variance, (0.5 + υ, 0.5+2υ, 0.5+3υ) = the three study precisions, ω_i = the study precision where the specific study is from).

k	τ^2	υ	ω_i	True I^2_{HT}	$I^2_{HT}(i)$(SE)	True I^2_T	$I^2_T(i)$(SE)	True I^2_S	$I^2_S(i)$(SE)
13	0	0.0	0.5	0	3.3 (0.5)	0	3.3 (0.5)	0	3.3 (0.5)
13	0	0.5	1.0	0	3.0 (0.5)	0	3.0 (0.5)	0	2.7 (0.5)
13	0	0.5	1.5	0	2.9 (0.5)	0	2.9 (0.5)	0	2.7 (0.5)
13	0	0.5	2.0	0	2.8 (0.5)	0	2.8 (0.5)	0	2.6 (0.5)
13	0	2.0	2.5	0	3.9 (0.5)	0	3.9 (0.5)	0	3.3 (0.5)
13	0	2.0	4.5	0	3.7 (0.5)	0	3.8 (0.5)	0	3.3 (0.5)
13	0	2.0	6.5	0	3.6 (0.5)	0	3.6 (0.5)	0	3.1 (0.5)
13	1	0.0	0.5	33.3	4.3 (0.7)	33.3	4.3 (0.7)	33.3	4.3 (0.7)
13	1	0.5	1.0	59.2	4.1 (0.7)	59.4	4.1 (0.7)	57.4	3.9 (0.7)
13	1	0.5	1.5	59.9	5.0 (0.7)	60.0	5.0 (0.7)	58.2	4.9 (0.6)
13	1	0.5	2.0	60.5	5.8 (0.6)	60.6	5.8 (0.6)	58.6	5.6 (0.6)
13	1	2.0	2.5	81.1	4.4 (0.4)	81.3	4.4 (0.4)	78.8	4.2 (0.4)
13	1	2.0	4.5	81.7	6.4 (0.4)	81.8	6.4 (0.4)	79.6	6.2 (0.4)
13	1	2.0	6.5	82.2	7.9 (0.4)	82.3	8.0 (0.4)	80.0	7.7 (0.4)
13	4	0.0	0.5	66.7	5.8 (0.6)	66.7	5.8 (0.6)	66.7	5.5 (0.6)
13	4	0.5	1.0	85.3	4.9 (0.3)	85.4	4.9 (0.3)	84.3	4.8 (0.3)
13	4	0.5	1.5	85.6	6.4 (0.3)	85.7	6.4 (0.3)	84.8	6.3 (0.3)
13	4	0.5	2.0	85.9	7.6 (0.3)	86.0	7.7 (0.3)	85.0	7.5 (0.3)
13	4	2.0	2.5	94.5	4.7 (0.1)	94.6	4.7 (0.1)	93.7	4.7 (0.1)
13	4	2.0	4.5	94.7	7.2 (0.1)	94.7	7.2 (0.1)	94.0	7.1 (0.1)
13	4	2.0	6.5	94.9	8.9 (0.1)	94.9	8.9 (0.1)	94.1	8.9 (0.1)
25	0	0.0	0.5	0	1.9 (0.4)	0	1.9 (0.4)	0	1.9 (0.4)
25	0	0.5	1.0	0	1.7 (0.4)	0	1.7 (0.4)	0	1.6 (0.4)
25	0	0.5	1.5	0	1.7 (0.4)	0	1.7 (0.4)	0	1.6 (0.4)
25	0	0.5	2.0	0	1.7 (0.4)	0	1.7 (0.4)	0	1.6 (0.4)
25	0	2.0	2.5	0	1.7 (0.4)	0	1.8 (0.4)	0	1.5 (0.3)
25	0	2.0	4.5	0	1.7 (0.4)	0	1.7 (0.4)	0	1.5 (0.3)

k	τ^2	υ	ω_i	True I^2_{HT}	I^2_{HT} (i)(SE)	True I^2_T	$I^2_T(i)$(SE)	True I^2_S	$I^2_S(i)$(SE)
25	0	2.0	6.5	0	1.7 (0.4)	0	1.7 (0.4)	0	1.5 (0.3)
25	1	0.0	0.5	33.3	2.5 (0.6)	33.33	2.5 (0.6)	33.33	2.5 (0.6)
25	1	0.5	1.0	59.6	2.1 (0.4)	59.7	2.1 (0.4)	57.7	2.0 (0.4)
25	1	0.5	1.5	60.0	2.6 (0.4)	60.0	2.6 (0.4)	58.1	2.6 (0.4)
25	1	0.5	2.0	60.2	3.2 (0.4)	60.3	3.2 (0.4)	58.4	3.1 (0.4)
25	1	2.0	2.5	81.5	2.2 (0.2)	81.6	2.2 (0.2)	79.1	2.2 (0.2)
25	1	2.0	4.5	81.7	3.5 (0.2)	81.8	3.5 (0.2)	79.5	3.4 (0.2)
25	1	2.0	6.5	82.0	4.5 (0.2)	82.1	4.5 (0.2)	79.7	4.4 (0.2)
25	4	0.0	0.5	66.7	3.3 (0.4)	66.7	3.3 (0.4)	66.7	3.3 (0.4)
25	4	0.5	1.0	85.5	2.3 (0.2)	85.5	2.3 (0.2)	84.5	2.3 (0.2)
25	4	0.5	1.5	85.7	3.2 (0.2)	85.7	3.2 (0.2)	84.7	3.2 (0.2)
25	4	0.5	2.0	85.8	4.0 (0.2)	85.9	4.0 (0.2)	84.8	4.0 (0.2)
25	4	2.0	2.5	94.6	2.3 (0.1)	94.6	2.3 (0.1)	93.8	2.3 (0.1)
25	4	2.0	4.5	94.7	3.8 (0.1)	94.7	3.8 (0.1)	94.0	3.8 (0.1)
25	4	2.0	6.5	94.8	5.0 (0.1)	94.8	5.0 (0.1)	94.0	5.0 (0.1)

Table 2

Reported summary statistics from six studies on CSF β-amyloid$_{42}$ (in pg/mL) as a function of ApoE4 Genotype (Author = the first author of the study, Year = the year of the publication, *n* = the sample size, SD = standard deviation)

Author	Year	*n*: ApoE4 +/−	Mean (SD): ApoE4 +	Mean (SD): ApoE4 −
Andreasen N [28]	1999	8/13	1641.00 (587.00)	1702.00 (339.00)
Jensen M [27]	1999	4/20	365.72 (85.79)	329.60 (139.97)
Tapiola T [29]	2000	13/25	500.00 (211.00)	522.00 (136.00)
Riemenschneide M [30]	2000	3/15	914.67 (11.37)	860.00 (194.00)
Sunderland T [26]	2004	57/85	389.00 (108.00)	443.00 (109.00)
Prince JA [31]	2004	32/86	697.00 (228.00)	840.00 (185.00)

Table 3

Study-specific indices of heterogeneity from six studies for estimating the mean difference of CSF β-amyloid$_{42}$ (in pg/mL) between ApoE4 genotypes

Author	Year	$I^2_{HT}(i)$	$I^2_T(i)$	$I^2_S(i)$	Leave-one-out estimate (in pg/mL) (95% CI)
Andreasen, N. [28]	1999	0.00	0.00	0.00	−30.87 (−134.68, 72.93)
Jensen, M. [27]	1999	0.13	0.15	0.05	−45.94 (−176.23, 84.35)
Tapiola, T. [29]	2000	0.00	0.00	0.00	−32.27 (−166.67, 102.12)
Riemenschneide, M. [30]	2000	0.27	0.31	0.10	−51.77 (−188.86, 85.33)
Sunderland, T. [26]	2004	0.00	0.00	0.00	−22.46 (−149.82, 104.91)
Prince, J.A. [31]	2004	0.34	0.40	0.12	−9.13 (−100.75, 82.4875)