Journal of Modern Applied Statistical Methods

Volume 7 | Issue 1 Article 24

5-1-2008

Measuring Overall Heterogeneity in Meta-Analyses: Application to CSF Biomarker Studies in Alzheimer's Disease

Chengjie Xiong
Washington University in St. Louis, chengjie@wubios.wustl.edu

Feng Gao
Washington University in St. Louis, eng@wubios.wustl.edu

Yan Yan
Washington University in St. Louis, yan@wubios.wustl.edu

Jingqin Luo
Washington University in St. Louis, rosy@wubios.wustl.edu

YunJu Sung
Washington University in St. Louis, yunju@wubios.wustl.edu

See next page for additional authors

Measuring Overall Heterogeneity in Meta-Analyses: Application to CSF Biomarker Studies in Alzheimer's Disease

Authors
Chengjie Xiong, Feng Gao, Yan Yan, Jingqin Luo, YunJu Sung, and Gang Shi

Journal of Modern Applied Statistical Methods
May, 2008, Vol. 7, No. 1, 286-296

1538 – 9472/08/$95.00

Measuring Overall Heterogeneity in Meta-Analyses: Application to CSF Biomarker Studies in Alzheimer's Disease

Chengjie Xiong[1], Feng Gao[1], Yan Yan[1,2], Jingqin Luo[1], YunJu Sung[1], & Gang Shi[1]
Washington University in St. Louis

The interpretations of statistical inferences from meta-analyses depend on the degree of heterogeneity in the meta-analyses. Several new indices of heterogeneity in meta-analyses are proposed, and assessed the variation/difference of these indices through a large simulation study. The proposed methods are applied to biomakers of Alzheimer's disease.

Keywords: Alzheimer's disease, heterogeneity, meta-analysis, standard errors, uncertainty interval.

Introduction

Medical practitioners and their patients make decisions within the context of a rapidly changing body of scientific evidence on medicine and health care system that influence the availability, accessibility, and cost of diagnostic tests and therapies (Sackett & 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 early 1990s (Antman, Kupelnick, Mosteller, & Chalmers, 1992), which compared the results of meta-analyses of trials of treatments for people who have suffered a heart attack as the trials were published with the recommendations of experts published in review articles and textbooks over the same time period. This showed a significant divergence between expert recommendations and the summaries of the trials.

Ineffective treatments were being recommended, and highly effective treatments were not. As a result, lives that could have been saved were lost, and resources were wasted. Systematic reviews can be very useful medical decision-making tools by objectively summarizing large amounts of information, identifying gaps in medical research and evidence, and identifying beneficial or harmful interventions. Clinicians can use systematic reviews to guide their patient care. Consumers and patients can use systematic reviews to help them make health care decisions. Policymakers can use systematic reviews to help them make decisions about the types of health care to provide.

Systematic reviews can provide convincing and reliable evidence relevant to many aspects of medical and biological research and health care (Egger & 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

[1]Division of Biostatistics, Washington University in St. Louis, St. Louis, MO 63110, U.S.A. [2] Department of Surgery, Washington University in St. Louis, St. Louis, MO 63110, U.S.A. E-mail format: name@wubios.wustl.edu: chengjie@; feng@; yan@; rosy@; yunju@; gang@. All authors have research interests in Biostatistics and Bioinformatics.

& Laird, 1986; Egger, Smith, & Phillips, 1997; & Whitehead & Whitehead, 1991).

Usually, each study is given a weight according to the precision of its results on summary statistics. Studies with good precisions 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 & Haenszel, 1959) or the Peto method (Yusuf, Peto, Lewis, Collins, & Sleight, 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, & 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. Fixed effects and random effects model for general continuous outcome and specific survival outcomes have also been described in Hedges and Olkin (1985), Earle & Wells (2000), Parmar, Torri, & Stewart, (1998) and Srinivasan & Zhou (1993).

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 meta-analysis commonly present a statistical test of heterogeneity among studies used in a 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 & Whitehead, 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 & Thompson, 2002).

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 a statistical test to characterize the degree of heterogeneity in a meta-analysis, Higgins and Thompson (2002) proposed a new measure of the extent of heterogeneity in a meta-analysis that overcomes 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 conclusion might be generalized to situations outside those investigated in the studies at hand. Their measure is easily interpretable by non-statisticians as the proportion of variation that was explained by the difference among studies. Further, the measure does 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 number of studies and types of outcome data.

In this article, several new indices are proposed that measure the heterogeneity from studies used in a meta-analysis. The proposed methodology can be regarded as a generalization of the index of heterogeneity proposed by Higgins and Thompson (2002). The difference among the proposed measures of heterogeneity are examined, along with the variation of each proposed measure when a large number of simulated meta-analyses are conducted. The proposed methodology is demonstrated by presenting an example to study possible cerebrospinal fluid (CSF) biomakers that could be used to identify subjects with high risk of developing Alzheimer's disease (AD) when they are still cognitively normal.

Indices of overall heterogeneity in a meta-analysis

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 $\hat{\sigma}_i^2$ be the associated estimated variance which is assumed to be known. Let $w_i = 1/\hat{\sigma}_i^2$ 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:

$$\hat{\theta} = \frac{\sum_{i=1}^{k} w_i \hat{\theta}_i}{\sum_{i=1}^{k} w_i}.$$

The variance of the summary estimate is given by

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

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

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

The estimated variance of the summary estimate is now given by

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

A test of homogeneity of the θ_i's is given by

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

which has a Chi-squared distribution with *k*-1 degrees of freedom under the assumption of homogeneity within the fixed effects model. Within the framework of the random effect model, a method of moment estimate to τ^2 can also be obtained as

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

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 term 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 definition of I^2 depends on the study specific estimates and is therefore based on the impact rather than the extent of heterogeneity in a meta-analysis. Second, the measure does not inherently depend on the number of studies in the meta-analysis. Third, the measure is not specific to a particular metric of treatment effect and therefore can be applied similarly irrespective of the type of outcome variables (e.g., dichotomous, continuous, and survival). Fourth, 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. For the latter, Higgins and Thompson (2002) used the following estimator

$$\hat{\sigma}_{HT}{}^2 = \frac{(k-1)\sum_{i=1}^{k} w_i}{(\sum_{i=1}^{k} w_i)^2 - \sum_{i=1}^{k} w_i^2}.$$

This, along with the method of moment estimator $\hat{\tau}^2$, results in the index of overall heterogeneity

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

Higgins and Thompson's (2002) intuitive conceptualization of the measure of heterogeneity is followed, and several new measures of heterogeneity are proposed. First, as pointed out by Takkouche et al. (Takkouche, Cadarso Surez, & Spiegelman, 1999), the typical within-study variance σ^2 can also be estimated by taking the reciprocal of the arithmetic mean weights:

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

Combine this with the method of moment estimator $\hat{\tau}^2$ in Equation (1) to obtain another index of overall heterogeneity

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

Another straightforward estimator to the typical within-study variance σ^2 can be obtained by simply averaging the within-study variances from all studies:

$$\hat{\sigma}_S{}^2 = \frac{\sum_{j=1}^{k} \frac{1}{w_j}}{k},$$

which, again along with the method of moment estimator $\hat{\tau}^2$ in Equation (1), results in another index of overall heterogeneity

$$I_S{}^2 = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}_S{}^2} = \frac{Q-(k-1)}{Q-k+1+\sum_{j=1}^{k}\frac{1}{w_j}\dfrac{\left[\left(\sum_{j=1}^{k} w_j\right)^2 - \sum_{j=1}^{k} w_j^2\right]}{k\sum_{j=1}^{k} w_j}}. \quad (4)$$

These proposed indices of heterogeneity are set to 0 if $Q \le (k-1)$. By Schwartz's inequality (Nobel & 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

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

Notice that for all these indices of overall heterogeneity, although they have different denominators, they share the same numerator, which is $Q-(k-1)$. If all within-study variations are exactly the same, then $I_{HT}^2 = I_T^2 = I_S^2$. Notice that the denominator of all these proposed overall 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 a study) and the between-study variation (i.e., heterogeneity).

Variation of the proposed overall measures of heterogeneity

Higgins and Thompson (2002) proposed several ways of estimating the variation associated with ${I_{HT}}^2$. They recommended the use of an uncertainty interval based on the statistical significance of Q due to the appropriate nominal coverage in their simulation studies. Because the other measures of overall heterogeneity we proposed here also depend on Q, we use similar test-based methods (Miettinen, 1976) to estimate the variability associated with ${I_T}^2$ and ${I_S}^2$ as well. More specifically, let

$$Z = \sqrt{2Q} - \sqrt{2k-3}.$$

Based on a well known normal approximation to Chi-squared distributions (Abramowitz & Stegun, 1965), when the degrees of freedom are large, Z follows approximately a standard normal distribution. Therefore, if $\ln(Q)$ is assumed a normal distribution, by equating Z with $\frac{\ln(Q)-\ln(k-1)}{SE(\ln(Q))}$, one can approximate the standard error of $\ln(Q)$ by

$$SE(\ln(Q)) = \frac{\ln(Q)-\ln(k-1)}{\sqrt{2Q}-\sqrt{2k-3}}.$$

This then results in a 95% uncertainty interval to ${I_T}^2$ as $[I_{T1}^2, I_{T2}^2]$, where

$${I_{T1}}^2 = \frac{\exp(\ln(Q)-1.96SE(\ln(Q)))-(k-1)}{\exp(\ln(Q)-1.96SE(\ln(Q)))+1-\frac{k\sum_{j=1}^{k} w_j^2}{\left(\sum_{j=1}^{k} w_j\right)^2}}$$

and

$${I_{T2}}^2 = \frac{\exp(\ln(Q)+1.96SE(\ln(Q)))-(k-1)}{\exp(\ln(Q)+1.96SE(\ln(Q)))+1-\frac{k\sum_{j=1}^{k} w_j^2}{\left(\sum_{j=1}^{k} w_j\right)^2}}.$$

Similarly, a 95% uncertainty interval to ${I_S}^2$ is $[I_{S1}^2, I_{S2}^2]$, where

$$I_{S1}^{\ 2} = \frac{\exp(\ln(Q) - 1.96SE(\ln(Q))) - (k-1)}{\exp(\ln(Q) - 1.96SE(\ln(Q))) - k + 1 + \sum_{j=1}^{k} \frac{1}{w_j} \frac{\left[\left(\sum_{j=1}^{k} w_j\right)^2 - \sum_{j=1}^{k} w_j^2\right]}{k\sum_{j=1}^{k} w_j}}$$

and

$$I_{S2}^{\ 2} = \frac{\exp(\ln(Q) + 1.96SE(\ln(Q))) - (k-1)}{\exp(\ln(Q) + 1.96SE(\ln(Q))) - k + 1 + \sum_{j=1}^{k} \frac{1}{w_j} \frac{\left[\left(\sum_{j=1}^{k} w_j\right)^2 - \sum_{j=1}^{k} w_j^2\right]}{k\sum_{j=1}^{k} w_j}}.$$

Comparison of the proposed overall indices of heterogeneity

Although mathematically, $I_S^2 \le I_T^2$, $I_{HT}^2 \le I_T^2$, it is important to understand how different these measures are when they are used to measure the overall 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, i.e., when all w_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.

A simulation study is performed to look at the performance of our proposed measures of overall and study-specific heterogeneity. We first examined the distributions and consistency of three different measures of overall heterogeneity, I_S^2, I_T^2, and I_{HT}^2, over a large number of simulated meta-analyses. Assume that the between-study variance is $\tau^2 = 4$. The number of studies in each simulated meta-analysis is $k = 3s + 1$ for s = 2, 4, 8, and 12. In each simulated meta-analysis, three different within-study variances are assumed such that the precision w_i is either 0.5+v, or 0.5+2v, or 0.5+3v for a range of v. More specifically, among $k = 3s+1$ studies in the meta-analysis, s+1 studies have within-study precision 0.5+v, and the other 2s studies are equally distributed with study precision 0.5+2v and 0.5+3v. A random effect model was used to generate the study-specific estimates such that the between-study component was generated from the normal distribution $N(5,4)$ through a linear transformation of the SAS Institute function RANNOR (SAS, 1999). One thousand independent simulated meta-analyses were performed such that study specific estimates from each individual simulation were independently generated. Table 1 presents the mean and the associated standard error for the three proposed measures of overall heterogeneity over 1000 simulated meta-analyses as a function of k and v (notice that parameter v here indicates a measure of difference among the study precisions).

From the simulated meta-analyses, it is clear that three different measures of overall heterogeneity are very consistent. In fact, under the assumption that the three measures of heterogeneity are estimating the same underlying trait of heterogeneity, the intraclass correlation coefficient (ICC) (Shrout & Fleiss, 1979) was computed over 1,000 simulated meta-analyses for each choice of *k* and *v*. All these computed ICCs were at least 0.99, indicating extremely high consistency among these measures.

Application to biomarker studies in Alzheimer 's disease

An application to the proposed overall measures of heterogeneity is presented to study possible biomakers that can be used to identify subjects with high risk of developing Alzheimer's disease (AD) when they are still cognitively normal. Researchers in Alzheimer's disease have identified Apolipoprotein E4 (ApoE4) alleles as a crucial genetic risk factor of AD (Myers, Schaefer, Wilson, et al., 1996). Although the pathological hallmarks of AD are the neurofibrillary tangles and the senile plaques in the brain (Braak & Braak, 1991, McKell, Price, Miller, Grant, Xiong, Berg, & Morris, 2004), the diagnosis of AD in living patients is still largely a clinical judgment based on careful neurological and/or neuropsychological examinations combined with results from other clinical tests.

Therefore, the search for biomarkers that could be used to diagnose AD from normal aging has been one of the primary research activities in AD. In several publications (Fagan, Roe, Xiong, et al., 2007, Sunderland, Linker, Mirza, et al., 2003), subjects with AD have been found to have decreased level of cerebrospinal fluid (CSF) β-amyloid$_{42}$ as compared to subjects with normal aging. Because AD is a progressive neurodegenerative disorder that leads to the death of brain cells that cannot be replaced once lost, it is important to assess the potential of these biomarkers to identify subjects that are at high risk of AD while they are still cognitively normal. The importance of such biomarkers is further highlighted by the fact that no pharmaceutical treatments are effective for the disease's later stages. Thus, whether CSF β-amyloid$_{42}$ is decreased among subjects of normal aging who are ApoE4 positive as compared to these who are ApoE4 negative is studied.

Although many publications have compared CSF β-amyloid$_{42}$ level between subjects with AD and these with normal aging (Fagan, Roe, Xiong, et al., 2007, Sunderland, Linker, Mirza 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 6 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 subjects who were not demented (Prince, Zetterberg, Andreasen, et al. 2004, Sunderland, Mirza, Putnam, et al., 2004, Jensen, Schroder, Blomberg et al., 1999, Andreasen, Hesse, Davidson et al., 1999, Tapiola, Pirtitla, Mehta, et al., 2000, Riemenschneider, Schmolke, Lautenschalager, et al, 2000). The summary statistics reported from these six published studies are presented in Table 2 (summary statistics from study by Prince, Zetterberg, Andreasen, et al., 2004) was obtained through eye-balling because only a graphical presentation on summary statistics was available in the publication).

Based on the proposed methodology and a random effect model, the pooled estimate to the mean difference of CSF β-amyloid$_{42}$ between subjects of normal aging who are ApoE4 positive and subjects 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. The observed significance level for the observed mean difference is 0.407. 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, Thompson, Deeks et al., 2003). Further, an estimated 95%

Table 1. Three Measures of Overall Heterogeneity from 1000 Simulated Meta-analyses
(k = the number of studies in meta-analyses,
(0.5+v, 0.5+2v, 0.5+3v) = the three within-study precisions)

k	v	Mean I^2_{HT} (SE) (%)	Mean I^2_T (SE) (%)	Mean I^2_S (SE) (%)
7	0	39.29 (0.85)	39.29 (0.85)	39.29 (0.85)
7	0.5	73.83 (0.58)	74.06 (0.57)	72.68 (0.58)
7	1.0	83.73 (0.45)	83.97 (0.45)	82.46 (0.46)
7	1.5	88.48 (0.31)	88.70 (0.30)	87.28 (0.32)
7	2.0	90.17 (0.29)	90.38 (0.29)	89.04 (0.31)
13	0	43.04 (0.66)	43.04 (0.66)	43.04 (0.66)
13	0.5	79.33 (0.34)	79.43 (0.34)	78.14 (0.35)
13	1.0	87.27 (0.19)	87.37 (0.19)	85.99 (0.21)
13	1.5	90.80 (0.15)	90.89 (0.15)	89.65 (0.16)
13	2.0	92.56 (0.14)	92.64 (0.14)	91.53 (0.15)
25	0.0	46.16 (0.50)	46.16 (0.50)	46.16 (0.50)
25	0.5	81.25 (0.18)	81.30 (0.18)	80.05 (0.19)
25	1.0	88.62 (0.13)	88.66 (0.12)	87.40 (0.14)
25	1.5	91.89 90.08)	91.93 (0.08)	90.81 (0.09)
25	2.0	93.69 (0.07)	93.73 (0.07)	92.76 (0.08)
37	0.0	47.41 (0.39)	47.41 (0.39)	47.41 (0.39)
37	0.5	81.88 (0.15)	91.91 (0.15)	80.69 (0.16)
37	1.0	89.22 (0.09)	89.25 (0.09)	88.03 (0.10)
37	1.5	92.16 (0.07)	92.18 (0.07)	91.10 (0.08)
37	2.0	93.94 (0.05)	93.96 (0.05)	93.03 (0.06)

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 of subjects who are ApoE4 positive,
n_- = the sample size of subjects who are ApoE4 negative,
Mean (SD)+ = mean (standard deviation) in subjects who are ApoE4 positive,
Mean (SD)- = mean (standard deviation) in subjects who are ApoE4 negative)

Author	Year	n_+ / n_-	Mean (SD)+	Mean (SD)-
Andreasen N	1999	8/13	1641.00 (587.00)	1702.00 (339.00)
Jensen M	1999	4/20	365.72 (85.79)	329.60 (139.97)
Tapiola T	2000	13/25	500.00 (211.00)	522.00 (136.00)
Riemenschneide M	2000	3/15	914.67 (11.37)	860.00 (194.00)
Sunderland T	2004	57/85	389.00 (108.00)	443.00 (109.00)
Prince JA	2004	32/86	697.00 (228.00)	840.00 (185.00)

uncertainty interval for I^2_{HT} is from 0.00 to 0.82, an estimated 95% uncertainty interval for I^2_T is from 0.00 to 0.88, and an estimated uncertainty interval for I^2_S is from 0.00 to 0.48 (the uncertainty intervals were truncated to 0 when the left limits were negative as similarly recommended in Higgins & Thompson, 2002).

If the heterogeneity is ignored in the meta-analysis, i.e., the between-study variance τ^2 is treated 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 subjects of normal aging who are ApoE4 positive and subjects 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 subjects of normal aging who are ApoE4 positive and subjects who are ApoE4 negative. This discrepancy on the statistical inference between the two approaches that either take into account of heterogeneity (i.e., random effect models) or ignore the heterogeneity (i.e., fixed effect models) further highlights the importance to assess the heterogeneity in meta-analyses.

Conclusion

We proposed several new indices that measure the overall heterogeneity among studies used in a meta-analysis. By estimating the typical within-study precisions, we developed indices that measure the degree of heterogeneity among studies by their impact to the overall conclusion of the meta-analysis. The proposed methodology can be regarded as a generalization of the index of heterogeneity proposed by Higgins and Thompson (2004). We assessed the variation associated with each proposed index of heterogeneity through a large simulation study of 1000 meta-analyses for a range of relevant parameters. We also examined the difference among the proposed overall measures of heterogeneity when a large number of meta-analyses were conducted. We found that these different indices provided highly consistent results in measuring the overall heterogeneity in meta-analyses. Finally, we demonstrated our proposed methodology by presenting an example to study possible biomakers that could

be used to identify subjects with high risk of developing Alzheimer's disease (AD) when they are still cognitively normal. The inconsistent statistical inferences to this real world example based on statistical approaches with or without taking into account of heterogeneity highlight the crucial role heterogeneity plays in meta-analyses.

References

Abramowitz, M., & Stegun, I. A. (1965). *Handbook of mathematical functions with formulas, graphs, and mathematical tables.* NY: Dover.

Andreasen, N., Hesse, C., Davidsson, P., et al. (1999). 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 in Neurology*, *56*, 673-680.

Antman, E. M., Lau, J., Kupelnick, B., Mosteller, F., & Chalmers, T. C. (1992). 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*, *268*, 240-248.

Braak, H., & Braak, E. (1991). Neuropathologic staging of Alzheimer-related changes. *Acta Neuropathologica,* 82, 239-259.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, *10*, 101-129.

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

Earle, C. C., & Wells, G. A. (2000). An assessment of methods to combine published survival curves. *Medical Decision Making, 20*, 104-111.

Egger, M., & Davey Smith, G. (1997). Meta-analysis: potentials and promise. *British Medical Journal*, *315*, 1371-1374.

Egger, M., Davey Smith, G., & Phillips, A. N. (1997). Meta-analysis; principles and procedures. *British Medical Journal*, *315*, 1533-1537.

Fagan, A. M., Roe, C. M., Xiong, C., et al. (2007). Cerebrospinal fluid tau/β-amyloid$_{42}$ ratio as a prediction of cognitive decline in nondemented older adults. *Archives in Neurology*,*64*. (doi:10.1001/archneur.64.3.noc60123).

Hardy, R. J., & Thompson, S. G. (1998). Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine*, *17*, 841-856.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando: Academic Press.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*, 1539-1558.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., et al. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327*, 557-560.

Jensen, M., Schroder, J., Blomberg, M., et al. (1999). Cerebrospinal fluid $A\beta 42$ is increased early in sporadic Alzheimer's disease and declines with disease progression. *Ann Neurol*, *45*, 504-511.

Laird, N. M., & Mosteller, F. (199). Some statistical methods for combining experimental results. *The International Journal of Technology Assessment in Health Care, 6*, 5-30.

Mantel, N., Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*., *22*, 719-748.

McKeel, D. W., Jr., Price, J. L., Miller, J. P., Grant, E. A., Xiong, C., Berg, L., & Morris, J. C. (2004). Neuropathologic criteria for diagnosing Alzheimer disease in persons with pure dementia of Alzheimer type. *Journal of Neuropathology and Experimental Neuropathology, 63*(10), 1028-1037.

Miettinen, O. (1976) Estimability and estimation in case-referent studies. *American Journal of Epidemiology*, *103*, 226-235.

Myers, R. H., Schaefer, E. J., Wilson, P. W. F., et al. (1996). Apolipoprotein E $\varepsilon 4$ association with dementia in a population-based study: The Framingham study. *Neurology*, *46*, 673-677.

Noble, B., & Daniel, J. W. (1977). *Applied linear algebra*. Englewood Cliffs, NJ: Prentice-Hall.

Parmar, M. K. B., Torri, V., & Stewart, L. A. (1998). Extracting summary statistics to perform meta-analyses of published literature for survival endpoints. *Statistics in Medicine, 17*, 2815-2834.

Prince, J. A., Zetterberg, H., Andreasen, N., et al. (2004). APOE ε 4 allele is associated with reduced cerebrospinal fluid levels of Aβ42. *Neurology*, *62*, 2116-2118.

Riemenschneider, M., Schmolke, M., Lautenschlager, N., et al. (2000). Cerebrospinal beta-amyloid $_{(1–42)}$ in early Alzheimer's disease: association with apolipoprotein E genotype and cognitive decline. *Neuroscience Letters*, *284*, 85-88.

Sackett, D. L., & Haynesm R. B. (1995). On the need for evidence-based medicine. *Evidence-Based Medicine*, *1*, 5-6.SAS Institute, Inc. (199).

SAS language (Ver. 8). Cary, NC

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86,* 420-428.

Srinivasan, C., & Zhou, M. (1993). A note on pooling Kaplan-Meier estimators. *Biometrics*, *49*, 861-864.

Sunderland, T., Linker, G., Mirza, N., et al. (2003). Decreased β-amyloid $_{1–42}$ and increased tau levels in cerebrospinal fluid of patients with Alzheimer's disease. *JAMA*, *289*, 2094-2103.

Sunderland, T., Mirza, N., Putnam, K. T., et al. (2004). Cerebrospinal fluid β-amyloid $_{1–42}$ and tau in control subjects at risk for Alzheimer's disease: The effect of apoE $\varepsilon 4$ allele. Biological Psychiatry, *56*, 670-676.

Takkouche, B., CadarsoSurez, C., Spiegelman. D. (1999). Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. *American Journal of Epidemiology*, 150, 206-215.

Tapiola, T., Pirttila, T., Mehta, P. D., et al. (2000). 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.

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

Yusuf, S., Peto, R., Lewis, J., Collins, R., & Sleight, P. (1985). Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Progress in Cardiovascular Diseases*. *27*, 335-371.