von Hippel *BMC Medical Research Methodology* (2015) 15:35
DOI 10.1186/s12874-015-0024-z

BMC
Medical Research Methodology

RESEARCH ARTICLE Open Access

# The heterogeneity statistic $I^2$ can be biased in small meta-analyses

Paul T von Hippel

## Abstract

**Background:** Estimated effects vary across studies, partly because of random sampling error and partly because of heterogeneity. In meta-analysis, the fraction of variance that is due to heterogeneity is estimated by the statistic $I^2$. We calculate the bias of $I^2$, focusing on the situation where the number of studies in the meta-analysis is small. Small meta-analyses are common; in the Cochrane Library, the median number of studies per meta-analysis is 7 or fewer.

**Methods:** We use Mathematica software to calculate the expectation and bias of $I^2$.

**Results:** $I^2$ has a substantial bias when the number of studies is small. The bias is positive when the true fraction of heterogeneity is small, but the bias is typically negative when the true fraction of heterogeneity is large. For example, with 7 studies and no true heterogeneity, $I^2$ will overestimate heterogeneity by an average of 12 percentage points, but with 7 studies and 80 percent true heterogeneity, $I^2$ can underestimate heterogeneity by an average of 28 percentage points. Biases of 12–28 percentage points are not trivial when one considers that, in the Cochrane Library, the median $I^2$ estimate is 21 percent.

**Conclusions:** The point estimate $I^2$ should be interpreted cautiously when a meta-analysis has few studies. In small meta-analyses, confidence intervals should supplement or replace the biased point estimate $I^2$.

**Keywords:** Meta-analysis, Heterogeneity, Bias

## Background

When different studies estimate the effect of a treatment or exposure, the estimates will vary from one study to another. Some of this between-study variance comes from random sampling error, while some may come from *heterogeneity*. There are several sources of heterogeneity, including differences in the treatment, the treated population, the study design, or the data analysis method. When there is no heterogeneity, estimates are said to be *homogeneous* and differ only because of random sampling error.

Heterogeneity is very important. If the existing studies of a treatment are homogeneous, or nearly homogeneous, then there is some assurance that the treatment will have a similar effect when applied to new subjects. On the other hand, if the existing studies are very heterogeneous, then unless the reasons for heterogeneity are well understood, the effect of the treatment on new subjects will be hard to predict [1].

Unfortunately, when studies are compared in a meta-analysis, it is often difficult to say anything definitive about heterogeneity. The reason for this difficulty is that most meta-analyses are small. One summary of the Cochrane Library reported that the median number of studies per meta-analysis was 7 [2], another summary reported that the median was 6 [3], and another reported that the median was just 3 [3]. With so few studies, the classical test for heterogeneity, Cochran's $Q$ [4], is not very informative because its result is as much a function of the number of studies as it is of the amount of heterogeneity. When the number of studies is large, $Q$ will often reject the null hypothesis even if the true extent of heterogeneity is trivial, but if the number of studies is small, $Q$ provides little power to reject the null hypothesis of homogeneity even if substantial heterogeneity is present [5]. The power of $Q$ and other homogeneity tests is further reduced when the studies in the meta-analysis

Correspondence: paulvonhippel.utaustin@gmail.com
Center for Health and Social Policy, LBJ School of Public Affairs, University of Texas, Austin, 2315 Red River, Box Y, Austin, TX 78712, USA

BioMed Central

© 2015 von Hippel; licensee BioMed Central. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

are unbalanced in size—for example, if one of the studies in the meta-analysis is much larger than the others [5].

To better describe heterogeneity, Higgins and Thompson [6] introduced the $I^2$ statistic, which was meant to improve in two ways on Cochran's $Q$. First, $I^2$ is more interpretable than $Q$; specifically, $I^2$ estimates the proportion of the variance in study estimates that is due to heterogeneity. Second, unlike $Q$, $I^2$ was meant to be independent of the number of studies; regardless of the number of studies, $I^2$ ranges from 0 to 1 because it estimates a proportion. The $I^2$ statistic is now used not just in meta-analysis but also in other analyses where we want to know what fraction of the variance in a set of estimates is due to heterogeneity [7-9].

$I^2$ does not eliminate the uncertainty that comes from having a small number of studies. No statistic can. In small meta-analyses, for the same reason that $Q$ has low power, $I^2$ is very imprecise. For example, if $Q$ fails to reject the null hypothesis of homogeneity, then the confidence interval around $I^2$ will usually include 0. In meta-analyses from the Cochrane Library, the 95% confidence interval around $I^2$ typically runs approximately from 0 to .60, implying that up to 60% of the between-study variance could be due to heterogeneity, or there could be no heterogeneity at all [2]. This is not a very informative conclusion. Unfortunately, the uncertainty of the $I^2$estimate is not obvious to the typical reader of a meta-analysis published in, for example, *Epidemiology* [10,11], the *American Journal of Epidemiology* [12,13], or the Cochrane Library [14]. These outlets do not report the confidence interval around $I^2$; they only report the point estimate $I^2$, which may give a false impression of precision.

In this note, we show that $I^2$ is not just imprecise; it is also biased. Depending on the circumstances, the bias of $I^2$ can be small or large, positive or negative, but the bias is largest when the number of studies is small and the true fraction of variance that is due to heterogeneity is either very large or very small. For example, in meta-analyses with 7 studies and no true heterogeneity, the $I^2$ statistic will on average lead us to believe that heterogeneity accounts for about 12% of the between-study variance. At the other extreme, with 7 studies and 80% of the variance due to heterogeneity, the $I^2$ statistic can on average lead us to believe that just 52% of the variance is due to heterogeneity. These biases of 12 to 28 percentage points are not trivial when one considers that, in the Cochrane Library, the median $I^2$ value is just 21% [2].

In the following sections, we calculate and illustrate the bias of $I^2$ and discuss implications for the statistics reported in meta-analyses.

## Methods

We use Mathematica software, version 8, to calculate the expectation and bias of $I^2$ analytically. This Methods section introduces notation, assumptions, and statistical properties, and describes the calculations that we submitted to Mathematica. The Results section will give the results of those calculations.

### Meta-analysis

Meta-analysis summarizes the results of $K$ studies, each of which has sample size $n_k$, $k = 1,\ldots,K$. In each study, there is a true effect $\beta_k$ estimated by $\hat{\beta}_k$, with a true standard error $\sigma_k$ estimated by $\hat{\sigma}_k$, or, equivalently, a true variance $\sigma_k^2$ estimated by $\hat{\sigma}_k^2$. With large $n_k$, the quantity $\left(\hat{\beta}_k - \beta_k\right)/\hat{\sigma}_k$ approaches a standard normal distribution according to the central limit theorem.

Two models can be used in meta-analysis: a *fixed-effects* model and a *random-effects* model. Some confusion is possible because the term fixed effects is used in two different senses [15]. In some literature, the term fixed effects means that the $K$ study effects $\beta_k$ are assumed to be homogeneous. We use the term fixed effects in its other sense, where it means that we seek only to generalize about the $K$ studies in the meta-analysis. The true effects $\beta_k$ can be either homogeneous or heterogeneous, but they are regarded as fixed quantities. Because of sampling error, the $K$ studies would produce different estimates $\hat{\beta}_k$ and $\hat{\sigma}_k$ if they were repeated, but the true effects $\beta_k$ and true standard errors $\sigma_k$ would not change.

Under a random-effects model, by contrast, we assume that the true effects $\beta_k$ in the meta-analysis were drawn at random from a larger population of effects, and we seek to make inferences about that larger population [16]. So the $\beta_k$ are not fixed quantities but random variables that would be different if a different sample were drawn from the population of effects.

### The estimand $\iota^2$

In order to understand the properties of the estimator $I^2$, we must first define the quantity that is being estimated. We call the estimand $\iota^2$. It represents the fraction of variance in the estimated effects $\hat{\beta}_k$ that is due to heterogeneity rather than measurement error.

More formally, the $\hat{\beta}_k$ vary from one study to another. The variance in $\hat{\beta}_k$ is partly due to the heterogeneity of the true effects $\beta_k$ and partly due to estimation error summarized by the standard errors $\sigma_k$. By the law of total variance we have

$$\begin{aligned} V\left(\hat{\beta}_k\right) &= V\left(\beta_k\right) + E\left(\sigma_k^2\right) \\ &= \tau^2 + \sigma^2 \end{aligned} \tag{1}$$

where $\tau^2 = V(\beta_k)$ is the heterogeneity variance or between-study variance, and $\sigma^2 = E\left(\sigma_k^2\right)$ is the average within-study variance. Under a fixed-effects model these

variances and expectations refer only to the $K$ effects $\beta_k$ and standard errors $\sigma_k$ in the meta-analysis. Under a random effects model $\tau^2$ refers to the larger population of effects, but $\sigma^2$ still refers only to the $K$ standard errors $\sigma_k$ in the meta-analysis, unless we are willing to regard the $\sigma_k$ as well as the $\beta_k$ as samples from a larger population.

The fraction of variance that is due to heterogeneity is

$$\iota^2 = \frac{V(\beta_k)}{V\left(\hat{\beta}_k\right)} = \frac{\tau^2}{\tau^2 + \sigma^2} \tag{2}$$

If $\iota^2 = 0$ then the effects $\beta_k$ are homogeneous; if $\iota^2 > 0$ then they are heterogeneous.

Note that, unlike some past definitions [6], our definition of $\iota^2$ does not assume equal standard errors $\sigma_1 = \sigma_2 = \ldots = \sigma_K$. Note also that $\iota^2$ is not an absolute measure of heterogeneity. Instead, $\tau^2$ is an absolute measure of heterogeneity, while $\iota^2$ compares $\tau^2$ to $\sigma^2$. When the estimation error is small, as it is if $n_k$ is large, then $\iota^2$ can be large even if $\tau^2$ is small [17].

### The naïve estimator $\hat{\iota}^2$

To estimate the fraction $\iota^2$, Higgins and Thompson [6] first derived the naïve estimator

$$\hat{\iota}^2 = 1 - \frac{df}{Q} \tag{3}$$

where $df = K-1$, $Q$ is Cochran's $Q$ statistic [4]

$$Q = \sum_{k=1}^{K} \frac{\left(\hat{\beta}_k - \hat{\bar{\beta}}\right)^2}{\hat{\sigma}_k^2} \tag{4}$$

and

$$\hat{\bar{\beta}} = \frac{\sum_{k=1}^{K} \hat{\sigma}_k^{-2} \hat{\beta}_k}{\sum_{k=1}^{K} \hat{\sigma}_k^{-2}} \tag{5}$$

is the precision-weighted average of the estimated effects.

The distribution of $\hat{\iota}^2$ depends on the distribution of $Q$. Under homogeneity, with large $n_k$, $Q$ has a central chi-square distribution with $df$ degrees of freedom.

Under heterogeneity, the large-$n_k$ distribution of $Q$ depends on whether we regard the effects as fixed or random. Under a random-effects model, $Q$ is distributed like a weighted sum of $K-1$ central $\chi_1^2$ variables, where the weights are given by a matrix function of $\tau^2$ and $\sigma_k^2$ [18]. If we make the simplifying assumption that all the standard errors are equal ($\sigma_k = \sigma$) then the weights are all equal to $1 + \tau^2/\sigma^2$ [18] or, in our notation $(1 - \iota^2)^{-1}$, so that

$$X = \left(1-\iota^2\right)Q \tag{6}$$

has a central chi-square distribution with $df$ degrees of freedom [18]. As $\iota^2$ gets small, we converge toward the homogeneous situation where $Q$ itself has a central chi-square distribution with $df$ degrees of freedom.

Under a fixed-effects model, by contrast, $Q$ has a non-central chi-square distribution with $df$ degrees of freedom and a non-centrality parameter of [19]

$$\lambda = \sum_{k=1}^{K} \frac{\left(\beta_k - \bar{\beta}\right)^2}{\sigma_k^2} \tag{7}$$

where $\bar{\beta}$ is the precision-weighted mean of the true effects $\beta_k$. If we make the simplifying assumption that all the standard errors are equal ($\sigma_k = \sigma$) then the non-centrality parameter reduces to

$$\begin{aligned} \lambda &= \frac{1}{\sigma^2} \sum_{k=1}^{K} \left(\beta_k - \bar{\beta}\right)^2 \\ &= K \frac{\tau^2}{\sigma^2} \\ &= K \frac{\iota^2}{1-\iota^2} \end{aligned} \tag{8}$$

The last line shows that $\lambda$ is an increasing function of $\iota^2$ and that that $\lambda = 0$ if $\iota^2 = 0$. So again, as $\iota^2$ gets small, $Q$ converges toward the central chi-square distribution that it has under homogeneity.

### The truncated estimator $I^2$

A shortcoming of the naïve estimator $\hat{\iota}^2$ is that it can be negative even though the estimand $\iota^2$ cannot. Negative values of $\hat{\iota}^2$ occur whenever $Q < df$, which is not a rare event. Figure 1 shows the probability that $\hat{\iota}^2$ is negative when the effects are homogeneous. The probability decreases as $df$ increases, but the probability is always greater than 50%.

To avoid negative estimates, Higgins and Thompson [6] suggested rounding them up to zero. The rounded or truncated estimator

$$I^2 = \max\left(0, \hat{\iota}^2\right) \tag{9}$$

is the estimator that is widely used today. $I^2$ cannot be negative but can be zero. Values of $I^2 = 0$ occur in about one-quarter of published meta-analyses [20].

### Expectation and bias of the estimators

The expectation of the naïve estimator $\hat{\iota}^2$ is

$$E\left(\hat{\iota}^2\right) = 1 - df\, E\left(\frac{1}{Q}\right) \tag{10}$$

This is easily calculated in the homogeneous case, where $1/Q$ is an inverse chi-square variable whose expectation is $1/(df - 2)$. It is just as easily calculated in

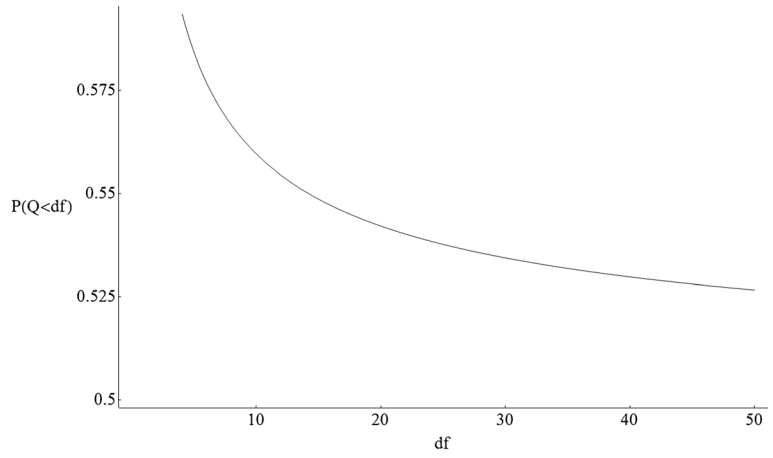


**Figure 1** The probability that a central chi-square variable $Q$ is less than its degrees of freedom.

the heterogeneous case with fixed effects; in that case, $1/Q$ is a scaled inverse chi-square variable with an expectation of $(1 - \iota^2)/(df - 2)$. The calculation is harder in the heterogeneous case with random effects; in that case, $1/Q$ is the scaled inverse of a noncentral chi-square variable. Although the expectation of this inverse has a closed-form solution [21], it is not transparent or easy to calculate by hand. However, we can calculate it using Mathematica.

The expectation of the truncated estimator $I^2$ is a little harder to calculate. It is the weighted average of two conditional expectations: the expectation of $I^2$ when $I^2 = 0$ and the expectation of $I^2$ when $I^2 > 0$. The probability that $I^2 = 0$ is $P(Q < df)$, and the probability that $I^2 > 0$ is $P(Q > df)$. Therefore the expectation of $I^2$ is

$$\begin{aligned} E(I^2) &= P(Q < df) \times 0 + P(Q > df) \times E(I^2 | Q > df) \\ &= P(Q > df) \times E\left(1 - \frac{df}{Q} | Q > df\right) \end{aligned} \tag{11}$$

Under homogeneity, $Q$ has a central chi-square distribution and the expectation $E(I^2)$ has a closed-form solution which Mathematica can calculate.

Under heterogeneity, the expectation $E(I^2)$ depends on whether we regard the effects as fixed or random. If effects are random, then $X = (1 - \iota^2)Q$ has a central chi-square distribution. The probability that $I^2 = 0$ is $P(X < (1 - \iota^2)df)$, and the probability that $I^2 > 0$ is $P(X > (1 - \iota^2)df)$. Therefore the expectation of $I^2$ is

$$\begin{aligned} E(I^2) = &\ P\left(X > (1-\iota^2)df\right) \\ &\times E\left(1 - (1-\iota^2)\frac{df}{X} \middle| X > (1-\iota^2)df\right) \end{aligned} \tag{12}$$

which again has a closed-form solution which Mathematica can calculate.

If instead effects are fixed, then the expectation $E(I^2)$ in (11) has no closed-form solution. But the expectation for specific values of $\iota^2$ and $df$ can be calculated using numerical integration in Mathematica.

## Results and discussion

### Expectation and bias of $I^2$ under homogeneity

Under homogeneity, there are two sources of bias in $I^2$, one positive and one negative. The positive source is larger, so the net bias in $I^2$ is positive.

The first source of bias is negative bias in the naïve estimator $\hat{\iota}^2 = 1 - df/Q$. Since the estimand $\iota^2$ is zero, the bias of $\hat{\iota}^2$ is the expectation

$$Bias(\hat{\iota}^2) = E(\hat{\iota}^2) = \frac{-2}{df - 2} \tag{13}$$

which is negative, and larger if $df$ is small.

The second source of bias arises when $\hat{\iota}^2$ is truncated to yield $I^2 = \max(0, \hat{\iota}^2)$. Since truncation rounds negative values up to 0, the resulting truncation bias is positive. When $df$ is small, truncation is more common (Figure 1), so the truncation bias is more severe.

While this intuitive explanation is helpful, it does not tell us whether the positive and negative components combine to produce a net bias that is positive or negative, large or small. To answer that question, we evaluate the expectation $E(I^2)$ in (11), which is also the bias since the estimand is $\iota^2 = 0$. Mathematica gives the bias as

$$\begin{aligned} Bias(I^2) &= E(I^2) \\ &= \left(\frac{df}{df - 2}\right) \frac{\left(\frac{df}{2e}\right)^{df/2} - \Gamma\left(\frac{df}{2}, \frac{df}{2}\right)}{\Gamma\left(\frac{df}{2} + 1\right)} \end{aligned} \tag{14}$$

where $\Gamma(df/2 + 1)$ is the gamma function and $\Gamma(df/2, df/2)$

is the upper incomplete gamma function (which has two arguments).

It is hard to tell by inspecting (14) whether the bias is positive or negative, small or large. To visualize the answer, Figure 2 plots the expectation $E(I^2)$, which is also the $Bias(I^2)$, as a function of the number of studies $K = df + 1$. The bias is always positive, indicating that the positive truncation bias outweighs the negative bias in in $\hat{\imath}^2$. The bias shrinks at a decreasing rate as $K$ grows. With $K = 3$ studies (which is the median in one summary of the Cochrane Library [22]), the bias is undefined because $E(I^2)$ is only defined if $df > 2$. With $K = 7$ studies (which is the median in another summary of the Cochrane Library [2]), the bias is .12. With $K = 10$ studies, the bias is .11; with $K = 50$ studies the bias is .06.

### Expectation and bias of $I^2$ under heterogeneity

Under heterogeneity, the expectation $E(I^2)$ depends on whether we regard the effects as fixed or random.

#### *Random-effects model*

With random effects, there are still two sources of bias in $I^2$, one positive and one negative. But now the positive source can be either smaller or larger than the negative source, so that the overall bias can be either negative or positive.

The first source of bias is negative bias in the naïve estimator $\hat{\imath}^2$:

$$\begin{aligned} Bias(\hat{\imath}^2) &= E\left(1-\frac{df}{Q}\right)-\imath^2 \\ &= \frac{2\imath^2-2}{df-2} \end{aligned} \tag{15}$$

This bias is always negative since $0 \le \imath^2 < 1$. The bias is larger if $df$ is small.

The second source of bias arises when $\hat{\imath}^2$ is truncated to yield $I^2 = \max(0, \hat{\imath}^2)$. Since truncation rounds negative values up to 0, truncation yields a positive bias. The truncation bias is smaller if $df$ is large or $\imath^2$ is large. This is because the probability of truncation is a little smaller when $df$ is large, and a lot smaller when $\imath^2$ is large. (From (12) the probability of truncation is $P(X > (1 - \imath^2)\, df)$, where $X \sim \chi^2_{df}$).

Intuitively, when $\imath^2$ is small, we approach the homogeneous case where the bias in $I^2$ is positive because of truncation. However, when $\imath^2$ is large, truncation is less common and the bias in $I^2$ approaches the bias of $\hat{\imath}^2$, which is negative.

More formally, under a random-effects model, the expectation $E(I^2)$ in (12) has a solution which Mathematica gives as

$$E\left(I^2\right) = \frac{\left(\begin{array}{c} -2e^{\frac{1}{2}df(\imath^2-1)}(df(\imath^2-1)-2)-df(\imath^2-1) \\ \times(df\imath^2-2)E_{-\frac{df}{2}}\left(\frac{1}{2}(df-df\imath^2)\right) \end{array}\right)}{(df-2)dfE_{1-\frac{df}{2}}\left(-\frac{1}{2}df(\imath^2-1)\right)} \tag{16}$$

where expressions of the form $E_n(z)$ represent the exponential integral function.

The expectation in (16) is in closed form but is even less transparent than its predecessor in (14). It is not clear from inspection whether the bias is large or small, positive or negative.

To visualize $E(I^2)$, Figure 3 gives a graphics grid displaying 9 plots of $E(I^2)$ as a function of $K$, for $\imath^2$ values between .1 and .9. In each plot, a dotted line is drawn at

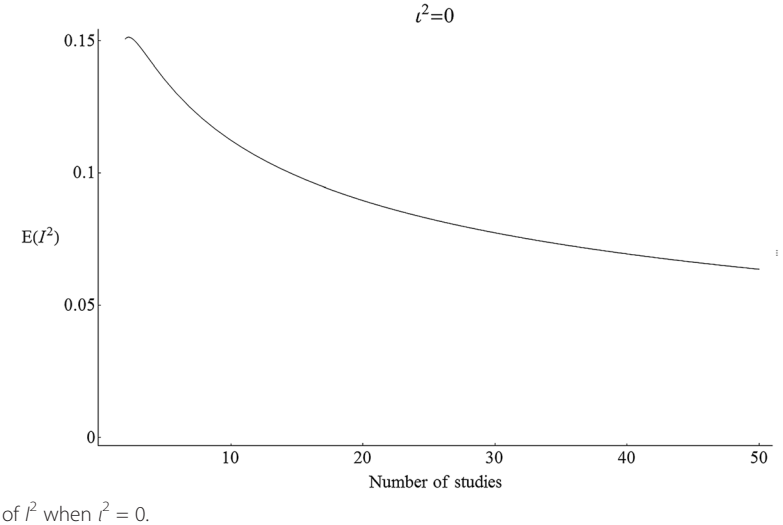


**Figure 2** The expectation of $I^2$ when $\imath^2 = 0$.

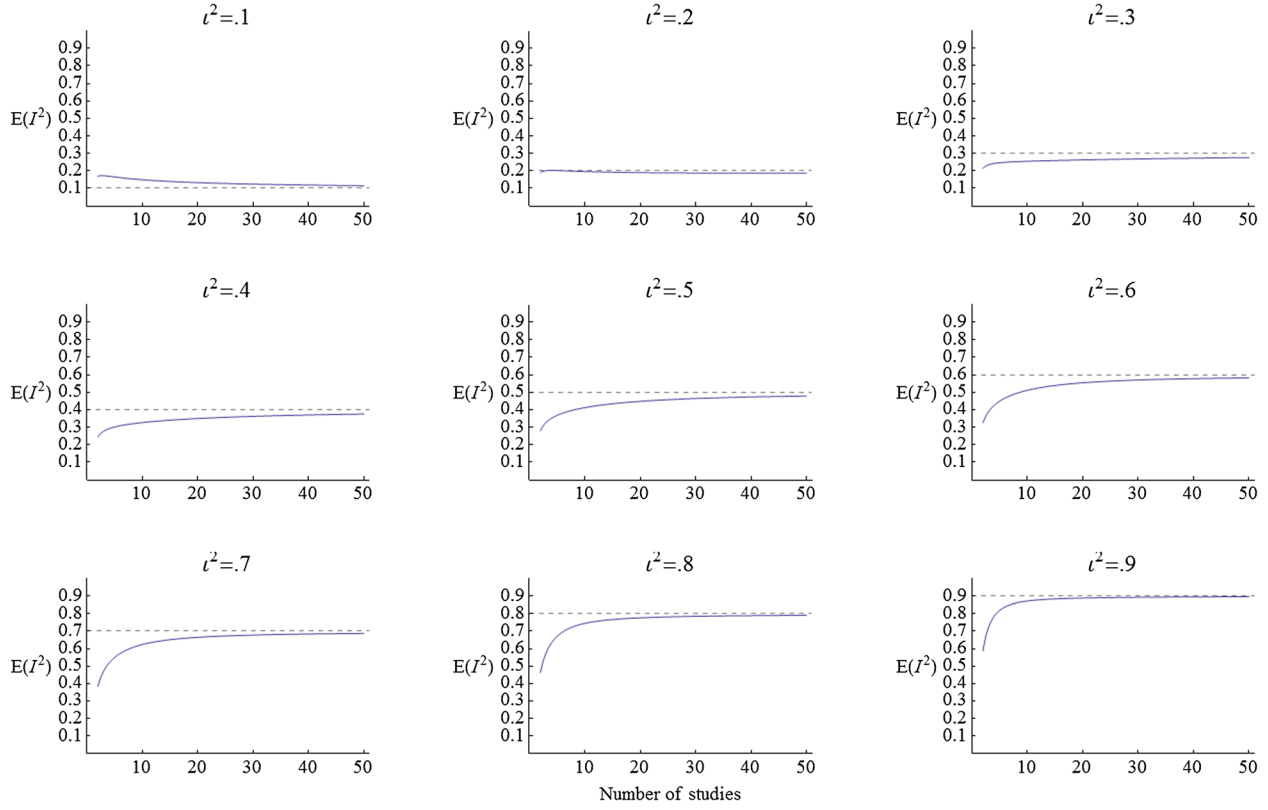


**Figure 3** The expectation of $I^2$ under a random-effects model, for $\iota^2$=.1,.2, …,.9.

the value of the estimand $\iota^2$, so that the bias of $I^2$ is the difference between the dotted line and the curve $E(I^2)$.

The bias is generally larger for small $K$. At $\iota^2$ = .1 the bias is positive. At $\iota^2$ = .2 there is practically no bias, and above $\iota^2$ = .2 the bias switches from positive to negative. As $\iota^2$ increases beyond .2 the bias gets larger for small $K$, but smaller for large $K$.

When $K$ is large there is practically no bias, particularly if $\iota^2$ is large as well. But when $K$ is small, as is often the case in meta-analysis, the bias can be noticeable even if $\iota^2$ is large. For example, if $\iota^2$ = .8 and $K$ = 7 (a typical or even high value for the Cochrane Library [2]), the expectation of $I^2$ is just .52.

### *Fixed-effects model*

Under heterogeneity with fixed effects, Mathematica gives the expectation of the naïve estimator $\hat{\iota}^2$ as

$$E\left(\hat{\iota}^2\right) = 1 + df\ 2^{df/2-2}\ \mathrm{e}^{-\lambda/2}(-1)^{-df/2}\lambda^{1-df/2} \times\left(\Gamma\left(\frac{df}{2}-1\right)-\Gamma\left(\frac{df}{2}-1,-\frac{\lambda}{2}\right)\right) \quad (17)$$

where $\lambda = K\iota^2/(1-\iota^2)$ from equation (8). However, this expression for $E(\hat{\iota}^2)$ is only real if $df$ is even.[a] If $df$ is odd, a much longer exact expression for $E(\hat{\iota}^2)$ can be derived using results in [21], or an approximation can be obtained numerically.

The bias of the naïve estimator is $\hat{\iota}^2-\iota^2$. Although it is not obvious from inspection, the bias is negative for $\iota$ < .8, and very slightly positive for $\iota \geq .8$.

The bias of the truncated estimator $I^2$ is a little different. Intuitively, when $\iota^2$ is small, we approach the homogeneous case where the bias in $I^2$ is positive because of truncation. However, as $\iota^2$ gets large, truncation is less common and the bias in $I^2$ approaches the bias of $\hat{\iota}^2$, which again is negative for $\iota$ < .8, and very slightly positive for $\iota \geq .8$.

The expectation of the truncated estimator $I^2$ can be calculated from equation (11) but under a fixed-effects model the solution no longer has a closed form, not even a complicated one. Instead, to evaluate $E(I^2)$ we use numerical integration in Mathematica.

Figure 4 is a graphics grid displaying 9 plots of $E(I^2)$ as a function of $K$, for $\iota^2$ values between .1 and .9. The bias is generally larger for small $K$. At $\iota^2$ = .1 the bias is positive. At $\iota^2$ = .2 there is practically no bias except for very small $K$. Above $\iota^2$ = .2 the bias switches from positive to negative. As $\iota^2$ increases from .3 to .5 the negative bias

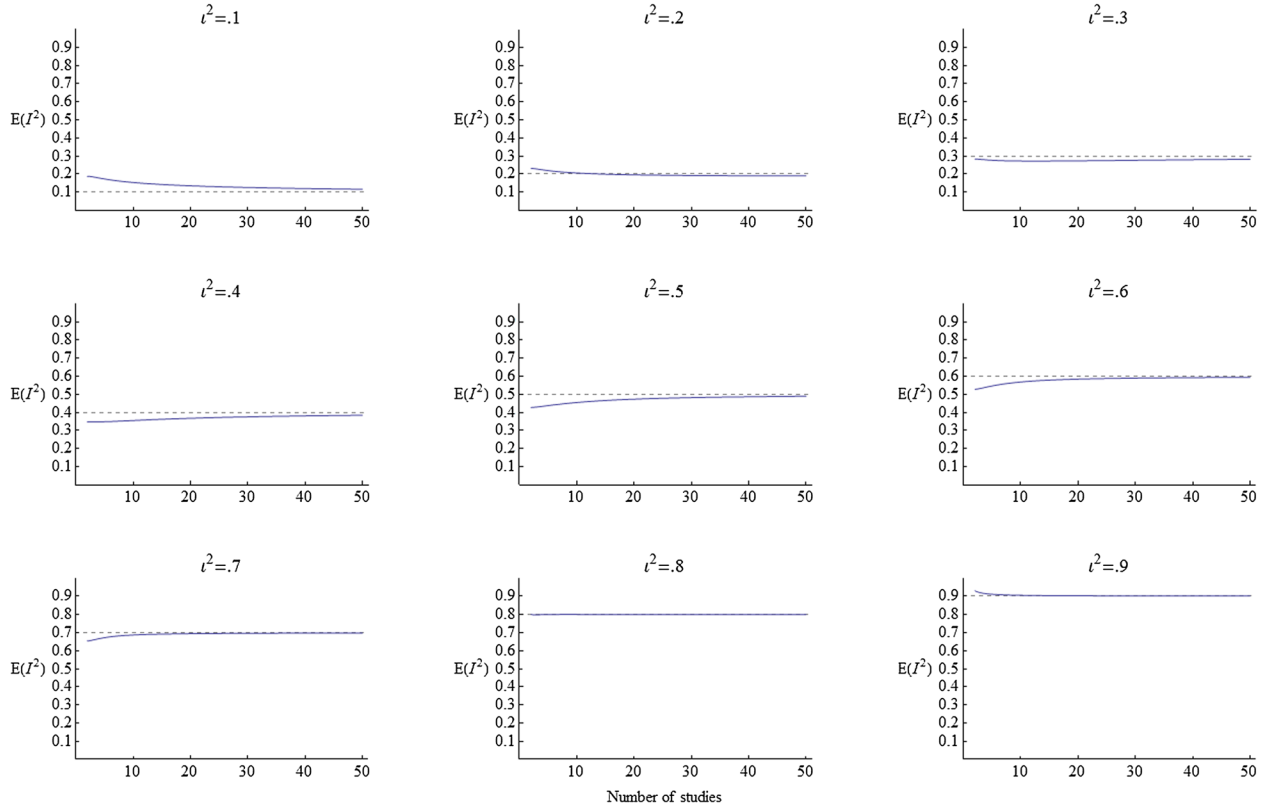


**Figure 4** The expectation of $I^2$ under a fixed-effects model, for $I^2$=.1,.2, …,.9.

gets larger, but as $\iota^2$ increases further from .6 to .7, the bias gets smaller and is increasingly restricted to small values of $K$, until at $\iota^2$ =.8 there is practically no bias. At $\iota^2$ =.9 the bias is positive again but very small and restricted to very small values of $K$.

In general, the bias is milder under the fixed-effects model than under the random-effects model, particularly if $\iota^2$ is large. For example, if $\iota^2$ = .8 and $K$ = 7 (a typical or even high value for the Cochrane Library [2]), the expectation of $I^2$ is just .52 under the random-effects model but is .80 (practically unbiased) under the fixed-effects model.

## Conclusions

We have shown that, in small meta-analyses, the widely used heterogeneity statistic $I^2$, which was already known to be imprecise, is biased as well. The bias shrinks as the number of studies $K$ grows, but since $K$ is often small in published meta-analyses, the bias of $I^2$ is often large in practice.

The bias and imprecision of $I^2$ are to some extent unavoidable and should not be taken as a criticism of the $I^2$ statistic itself. All statistics are imprecise in small samples, and any reasonable estimator of the heterogeneity fraction $\iota^2$ will be biased when the true value of $\iota^2$ is close to 0. The reason for the bias is fundamental. Like the estimand $\iota^2$, any reasonable estimator should be limited to nonnegative estimates, but the expectation of those nonnegative estimates will be positive and will exceed $\iota^2$ when the true value of $\iota^2$ is close to 0.

Similar bias has been observed in the heterogeneity variance $\tau^2$. Any reasonable estimator of $\tau^2$ will be limited to nonnegative values, and this will cause bias when the true value of $\tau^2$ is close to zero [15,23]. Estimators of $\tau^2$ have been constructed that are less biased or more precise under some circumstances, but all nonnegative estimators are biased when the true value of $\tau^2$ is close to zero [24].

Despite its bias and imprecision, the $I^2$ statistic remains useful. In large meta-analyses, $I^2$ can be precise with little bias, and even in small meta-analyses it is better to have a biased and imprecise estimate of $\iota^2$ than it is to have no estimate at all. In addition, although the bias of $I^2$ depends to some extent on the number of studies $K$, $I^2$ is much less dependent on $K$ than $Q$ is.

Nevertheless, $I^2$ should be presented and interpreted cautiously in small meta-analyses. Perhaps the most straightforward response to the bias and imprecision of $I^2$ is to report a 95% confidence interval in addition to—or even instead of—the point estimate $I^2$. Although methods for

calculating confidence intervals around $I^2$ can be a bit complicated [6,19,23,25], the best methods have good coverage and they give a sense of the range of possible $\imath^2$ values without highlighting a point estimate that may be biased and imprecise. While some meta-analyses do report confidence intervals around $I^2$ [26], such confidence intervals are not included in recent meta-analysis published in journals such as *Epidemiology* [10,11], the *American Journal of Epidemiology* [12,13], or the Cochrane Library. Journals publishing meta-analysis should consider requiring confidence intervals for $\imath^2$.

In small meta-analyses, confidence intervals for $\imath^2$ are often very wide [2] but their width tells us something. The width of the confidence intervals tells us how little information a small meta-analysis typically provides about heterogeneity. In many small meta-analyses, we may not be able to estimate heterogeneity with much precision; in fact, we may have little confidence in any estimate beyond the average effect size. No statistic can change the limitations of small meta-analyses, and the statistics that we report should make those limitations clear.

## Endnote

[a]We filed a bug report with Wolfram Research regarding Mathematica's failure to provide a real solution for odd *df*.

### Competing interests

The author declares that he has no competing interests.

### Acknowledgments

This article was not part of any funded project. I thank Erika Patall, James Pustejovsky, Tasha Beretvas, and journal reviewers for comments on earlier drafts.

Received: 13 December 2014 Accepted: 24 March 2015
Published online: 14 April 2015

### References

1. Melsen WG, Bootsma MCJ, Rovers MM, Bonten MJM. The effects of clinical and statistical heterogeneity on the predictive values of results from meta-analyses. Clin Microbiol Infect. 2014;20(2):123–9.
2. Ioannidis JPA, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analyses. BMJ. 2007;335(7626):914–6.
3. Davey J, Turner RM, Clarke MJ, Higgins JP. "Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: a cross-sectional, descriptive analysis". BMC Med Res Methodol. 2011;11(1):160.
4. Cochran WG. The combination of estimates from different experiments. Biometrics. 1954;10:101–29.
5. Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. Stat Med. 1998;17(8):841–56.
6. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;21(11):1539–58.
7. Koedel C. An empirical analysis of teacher spillover effects in secondary school. Econ Educ Rev. 2009;28(6):682–92.
8. Koedel C, Parsons E, Podgursky M, Ehle M. Teacher preparation programs and teacher quality: are there real differences across programs? Washington, DC: American Institutes for Research; 2012. CALDER working paper 63.
9. von Hippel PT, Osborne C, Lincove A, Bellows L, Mills N. The challenges of seeking exceptional teacher preparation programs among many noisy estimates. Rochester, NY: Social Science Research Network; 2014. SSRN Scholarly Paper ID 2506935.
10. Kivimäki M, Batty GD, Ferrie JE, Kawachi I. Cumulative meta-analysis of job strain and CHD. Epidemiology. 2014;25(3):464–5.
11. Aune D, Saugstad OD, Henriksen T, Tonstad S. Physical activity and the risk of preeclampsia: a systematic review and meta-analysis. Epidemiology. 2014;25(3):331–43.
12. Crippa A, Discacciati A, Larsson SC, Wolk A, Orsini N. Coffee consumption and mortality from all causes, cardiovascular disease, and cancer: a dose–response meta-analysis. Am J Epidemiol. 2014;180(8):763–75.
13. Kim Y, Je Y. Dietary fiber intake and total mortality: a meta-analysis of prospective cohort studies. Am J Epidemiol. 2014;180(6):565–73.
14. Cochrane Collaborative, "Cochrane Library", 2015. [Online]. Available: http://www.cochranelibrary.com/. [Accessed: 25-Feb-2015].
15. Hedges LV, Vevea JL. Fixed- and random-effects models in meta-analysis. Psychol Methods. 1998;3(4):486–504.
16. Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. J R Stat Soc A Stat Soc. 2009;172(1):137–59.
17. Rücker G, Schwarzer G, Carpenter JR, Schumacher M. "Undue reliance on I2 in assessing heterogeneity may mislead". BMC Med Res Methodol. 2008;8(1):79.
18. Biggerstaff BJ, Jackson D. The exact distribution of Cochran's heterogeneity statistic in one-way random effects meta-analysis. Stat Med. 2008;27(29):6093–110.
19. Hedges LV, Pigott TD. The power of statistical tests in meta-analysis. Psychol Methods. 2001;6(3):203–17.
20. Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. Stat Med. 2000;19(13):1707–28.
21. Bock ME, Judge GG, Yancey TA. A simple form for the inverse moments of non-central $\chi$2 andF random variables and certain confluent hypergeometric functions. J Econometrics. 1984;25(1–2):217–34.
22. Turner RM, Davey J, Clarke MJ, Thompson SG, Higgins JP. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews. Int J Epidemiol. 2012;41(3):818–27.
23. Viechtbauer W. Bias and efficiency of meta-analytic variance estimators in the random-effects model. J Educ Behav Stat. 2005;30(3):261–93.
24. Chung Y, Rabe-Hesketh S, Choi I-H. Avoiding zero between-study variance estimates in random-effects meta-analysis. Stat Med. 2013;32(23):4071–89.
25. Hartung J, Knapp G. On confidence intervals for the among-group variance in the one-way random effects model with unequal error variances. J Stat Plann Infer. 2005;127(1–2):157–77.
26. Ray KK, Seshasai SRK, Erqou S, Sever P, Jukema JW, Ford I, et al. Statins and all-cause mortality in high-risk primary prevention: a meta-analysis of 11 randomized controlled trials involving 65,229 participants. Arch Intern Med. 2010;170(12):1024–31.

**Submit your next manuscript to BioMed Central and take full advantage of:**

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central