BMJ Open

A plea for routinely presenting prediction intervals in meta-analysis

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2015-010247
Article Type:	Research
Date Submitted by the Author:	12-Oct-2015
Complete List of Authors:	IntHout, Joanna; Radboud university medical center, Radboud Institute for Health Sciences, Health Evidence Ioannidis, John; Stanford University, Stanford Prevention Research Center, Department of Medicine and Department of Health Research and Policy Rovers, Maroeska; The Radboud university medical center Goeman, Jelle; Radboud university medical center, Radboud Institute for Health Sciences, Health Evidence
Primary Subject Heading:	Research methods
Secondary Subject Heading:	Communication, Epidemiology, Evidence based practice
Keywords:	Meta-analysis, Prediction interval, Heterogeneity, Random effects, Clinical trial, Cochrane Database of Systematic Reviews

A plea for routinely presenting prediction intervals in meta-analysis

Authors: Joanna IntHout*, John PA Ioannidis, Maroeska M Rovers, Jelle J Goeman

Joanna IntHout[1*]
* Corresponding author
Email: Joanna.IntHout@radboudumc.nl

John PA Ioannidis[2,3,4,5]
Email: JIoannid@stanford.edu

Maroeska M Rovers[1]
Email: Maroeska.Rovers@radboudumc.nl

Jelle J Goeman[1]
Email: Jelle.Goeman@radboudumc.nl

[1] Radboud university medical center, Radboud Institute for Health Sciences (RIHS), Mailbox 133, P.O. box 9101, 6500 HB Nijmegen, The Netherlands

[2] Stanford Prevention Research Center, Department of Medicine, Stanford University School of Humanities and Sciences, Stanford, CA 94305, USA

[3] Department of Health Research and Policy, Stanford University School of Medicine, Stanford, CA 94305, USA

[4] Department of Statistics, Stanford University School of Humanities and Sciences, Stanford, CA 94305, USA

[5] Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA 94305, USA

ABSTRACT

Objectives: Evaluating the variation in the strength of the effect across studies is a key feature of meta-analyses. This variability is reflected by measures like τ^2 or I^2 but their clinical interpretation is not straightforward. A prediction interval is less complicated: it presents the expected range of true effects in similar studies. We aimed to show the advantages of having the prediction interval routinely reported in meta-analyses.

Design: We show how the prediction interval can help understand the uncertainty about whether an intervention works or not. To evaluate the implications of using this interval to interpret the results, we selected the first meta-analysis per intervention review of the Cochrane Database of Systematic Reviews Issues 2009-2013 with a dichotomous (n=2009) or continuous (n=1254) outcome, and generated 95% prediction intervals for them.

Results: In 72.4% of 479 statistically significant (random effects $p<0.05$) meta-analyses in the Cochrane Database 2009-2013 with heterogeneity ($I^2>0$), the 95% prediction interval suggested that the intervention effect could be null or even be in the opposite direction. In 20.3% of those 479 meta-analyses, the prediction interval showed that the effect could be completely opposite to the point estimate of the meta-analysis. We demonstrate also how the prediction interval can be used to calculate the probability that a new trial will show a negative effect and to improve the calculations of the power of a new trial.

Conclusions: The prediction interval reflects the variation in treatment effects over different settings, including what effect is to be expected in future patients such as the patients that a clinician is interested to treat. Prediction intervals should be routinely reported to allow more informative inferences in meta-analyses.

Word count: 274

STRENGTHS AND LIMITATIONS OF THIS STUDY

- In many meta-analyses there is large variation in the strength of the effect.
- The prediction interval helps in the clinical interpretation of the heterogeneity by estimating what true treatment effects can be expected in future settings.
- In case of heterogeneity, prediction intervals will show a wider range of expected treatment effects than confidence intervals, and thus may lead to different conclusions. This occurred in over 70% of statistically significant meta-analyses with heterogeneity of the Cochrane Database of Systematic Reviews. Completely opposite effects were not excluded in over 20% of those meta-analyses.
- Prediction intervals should be routinely reported to allow more informative inferences in meta-analyses.
- Limitations are that the calculations and inferences for the prediction interval are based on the normality assumption, which is difficult to ensure. Further, the interval will be imprecise if the estimates of the summary effect and the τ^2 are imprecise, for example if they are based on only a few, small studies. Inferences based on the prediction interval are only valid for settings that are similar (exchangeable) to those on which the meta-analysis is based.

INTRODUCTION

Interventions may have heterogeneous effects across studies because of differences in study populations, interventions, follow-up length, bias, and other factors.[1] Nevertheless, the usual reporting of a meta-analysis is focused on the summary effect size combined with a confidence interval (CI) and p-value. Typically also some measure of the between-study heterogeneity is presented such as τ^2 or the inconsistency measure I^2.[2,3] However, neither of these two metrics can readily point to the clinical implications of the observed heterogeneity. Our objective in the current article is to show the potential advantages of obtaining and reporting the prediction interval routinely in meta-analyses because its clinical meaning is much more straightforward. The prediction interval presents the heterogeneity in the same metric as the original effect size measure, in contrast to τ^2 or I^2. Reporting a prediction interval in addition to the summary estimate and confidence interval will illustrate which range of true effects can be expected in future settings. We describe its merits and provide working examples to show how it can be calculated.

METHODS

1. INTERPRETATION OF HETEROGENEITY

Between-study variation in the magnitude of treatment effects cannot be neglected. One of the main merits of a meta-analysis may even be that it reveals the variation of effects in different studies.[4] Therefore summarizing the findings of a meta-analysis in a single summary value sacrifices potentially informative variation.[5] However, the information that can be directly retrieved from τ^2 and I^2 with respect to the variation in the effects is limited. The clinical interpretation of I^2 is ambiguous: a high I^2 does not necessarily imply that the study effects are dispersed over a wide range[6] and a low I^2 might correspond to high dispersion[7], because I^2

depends on sample size. With very large (highly precise) studies, even tiny differences in effect size may result in a high I^2, while with small (imprecise) studies, very different treatment effects can yield an I^2 of 0. Dispersion in treatment effects is better reflected by τ because τ is the standard deviation of the between-study effects. One could for example estimate the ratio of the effect size over τ, which can convey how many times larger the treatment effect is compared to the standard deviation of the effect across studies.[8] But this may still be not very intuitive to a clinical reader. Another popular way to express variation in effect sizes is the CI, e.g. the 95% CI. The CI in a random effects model contains highly probable values for the summary treatment effect. However, it does not convey what range of treatment effects are likely to be seen in other patients, e.g. in the next study or in the patients a clinician wants to treat in her clinic.

2. PREDICTION INTERVALS

Not so often reported but much more insightful is the prediction interval.[9] A prediction interval always presents the heterogeneity on the same scale as the original outcomes, in contrast to τ, τ^2 or I^2. A 95% prediction interval estimates where the true effects are to be expected for 95% of similar (exchangeable) studies that might be conducted in the future.[4] Therefore it is well suited to evaluate the variability of the effect of an intervention over different settings. In the absence of between-study heterogeneity, the prediction interval coincides with the respective CI. However, in case of heterogeneity a prediction interval covers a wider range than a CI. Consequently, in case of a statistically significant effect (where all values of the 95% CI are on the same side of the null) the corresponding 95% prediction interval may indicate that values are possible on both sides of the null. This means that there will be settings where conclusions based on CIs will not hold. In the same framework, one can also calculate the probability that the true effect will be harmful (on the

other side of the null) in a next study. Table 1 presents an overview of measures of between-study heterogeneity.

3. EXAMPLE: TOPICAL STEROIDS FOR NASAL POLYPS

A 2012 review on the use of topical steroids for treatment of chronic rhinosinusitis with nasal polyps, based on seven randomized studies, resulted in a larger decrease in overall symptom scores in favor of steroids compared to placebo, reflected by a standardized mean difference (SMD) of -0.51, with a 95% CI from -0.96 to -0.07 (Figure 1).[10] The I^2 was 73.9% (95% CI, 44.2% to 87.8%), which can be considered substantial heterogeneity[11], and the estimated τ^2 was 0.148. Notwithstanding these numbers, it is difficult to evaluate what the clinical consequences of this heterogeneity may be for future settings.

In order to estimate the prediction interval for the SMD we need the point estimate of the SMD, its standard error (SE) and the estimated τ^2. We derive the SE from the 95% CI of the SMD (formula 1 appendix), which results in an SE of 0.182. We can calculate the standard deviation of the prediction interval SD_{PI} as $\sqrt{(0.148 + 0.182^2)}$ and the lower and upper limit of the 95% prediction interval as $-0.51 \pm 2.45 \times SD_{PI}$. The value 2.45 results from the $t_{0.05/2,6}$ distribution. Prediction intervals with a different coverage could be calculated by using a different t-value, e.g. $t_{0.20/2,6}$ for an 80% prediction interval (formula 1 appendix).

The resulting prediction interval, ranging from -1.55 to 0.53, can be interpreted as the 95% range of true SMDs to be expected in similar studies. We present it in Figure 1 as a rectangle below the diamond for the 95% CI.[12] The prediction interval contains values below zero, which corresponds to a decrease in symptom scores of at best approximately 1.5 SD after steroid use compared to placebo. But it also contains values above zero which means that the

steroids may exhibit no or even a harmful effect (SMD>0) in some settings, with a (95%) worst case increase in SMD of 0.53. Consequently, the effect in a new study may be even the exact opposite to the summary point estimate of the meta-analysis, i.e. an increase of 0.51 instead of a decrease of -0.51 may occur. The estimated probability that the true effect of the steroids will be null or higher in a new study is equal to 13.6%, based on the t-distribution with 6 degrees of freedom (formula 2 appendix).

Figure 1 approximately here

4. COCHRANE DATABASE

In order to investigate how often there is a discrepancy in conclusions based on prediction intervals and CIs we evaluated this in statistically significant meta-analyses ($p<0.05$ by random effects calculations) of the Cochrane Database of Systematic Reviews Issues 2009-2013, kindly provided by the UK Cochrane Editorial Unit. To avoid subjectivity in the selection we used the first meta-analysis with a dichotomous or continuous outcome and based on at least two studies in the Data and Analyses section. Details can be found in another paper.[13] In brief, of a total of 3263 meta-analyses, 920 were statistically significant: 479 with an estimated $I^2>0$ and 441 with an estimated $I^2=0$.

5. CALCULATIONS

We used the Hartung-Knapp/Sidik-Jonkman[14] random effects meta-analysis approach combined with the empirical Bayes estimator for τ^2. We estimated τ^2 for all meta-analyses, even when the authors originally performed a fixed effects analysis. Prediction intervals were calculated according to formula 1 (Appendix). We categorized the statistically significant meta-analyses with heterogeneity ($\tau^2>0$) by number of studies (2-6 studies or >6) and

heterogeneity (I^2<30%, 30-60%, or >60%, based on the Cochrane Handbook[11] stating that an I^2 between 30% and 60% corresponds to moderate heterogeneity). For significant meta-analyses where the heterogeneity estimate was zero, we assessed the impact of possibly low but non-zero heterogeneity by assuming an I^2 of 20%, calculating prediction intervals using formula 3 (appendix). Categorical outcomes were compared between groups by means of the chi-square test. We used R software[15] version 3.1.2 and the R packages metafor[16] version 1.9-5 and meta[17] version 4.1-0.

RESULTS

Overall, 132 (27.6%) of the 479 statistically significant meta-analyses with an $I^2 > 0$ had both the 95% CI and the 95% prediction interval excluding the null effect (Table 2). Consequently, almost three-quarter (72.4%) had a prediction interval that contained the null effect. This means that it is likely that for these comparisons some patient populations might experience null effects or effects in the opposite direction, i.e. a treatment might be more harmful than the comparator even though the point estimate suggests benefit (or vice versa). Not surprisingly, significant meta-analyses with low heterogeneity more often had prediction intervals that excluded the null than meta-analyses with high heterogeneity. The percentage of prediction intervals containing the null effect was slightly higher for meta-analyses with a continuous outcome (80.4%) than for those with a dichotomous outcome (65.8%) (p<0.001), but not significantly different for meta-analyses based on more than six studies (74.1%) than for those with at most six studies (69.1%) (p=0.25). (Table W1).

PREDICTION INTERVALS CONTAINING THE OPPOSITE EFFECT

If the prediction interval just includes the null effect, this may be less worrying than when it contains the exact opposite effect of the pooled summary effect, e.g. if it contains an OR of

0.5 when the meta-analysis summary estimate is an OR of 2, or if it contains an SMD of -0.7 when the summary estimate was 0.7. Of the 479 significant meta-analyses with an $I^2 > 0$, 97 (20.3%) had a prediction interval that contained the opposite effect. This percentage was higher for the meta-analyses with a continuous outcome (65/219, 29.7%) than for those with a dichotomous outcome (32/260, 12.3%) (p<0.001). It occurred also more frequently in meta-analyses with more than 6 primary studies (57/139, 41.0% and 30/178, 20.3% for meta-analyses with a continuous or dichotomous outcome, respectively) than for those based on at most 6 studies (8/80, 10.0% and 2/82, 2.4%) (p<0.001 and p=0.001, respectively).

Table 2 approximately here

META-ANALYSES WITH ESTIMATED I^2=0

A substantial part of meta-analyses have an estimated I^2 of 0. However, there is typically very large uncertainty about the exact amount of heterogeneity and this is demonstrated by very large 95% CIs for the values of I^2.[18] The same applies to τ: an estimate of 0 is often accompanied by large uncertainty. The true I^2 and τ are unlikely to ever be exactly 0, although low values are possible. To assess the impact of possibly low but non-zero heterogeneity among the 441 Cochrane meta-analyses with estimated I^2=0 and statistically significant results, we imputed an I^2=20% (suggestive of low between-study heterogeneity). Under this assumption, in 329 (74.6%) of these 441 meta-analyses the 95% prediction interval would span both sides of the null (Table 1), similar for meta-analyses with a dichotomous (74.7%) or continuous (74.4%) outcome (Table W1). This is a sensitivity analysis that is useful to perform to see whether the inferences of a meta-analysis that seemingly does not have detectable heterogeneity may be influenced by even a small amount of heterogeneity.

DISCUSSION AND OUTLOOK

In meta-analyses a CI is inadequate for clinical decision making because it only summarizes the average effect for the average study. The prediction interval is more informative as it shows the range of possible effects in relation to harm and clinical benefit thresholds. While we have focused on the situation where the separating threshold is the null, a different threshold may be considered. For example, in the prediction interval framework one can calculate the probability that an effect is larger than B, where B may be a clinically meaningful effect (if the treatment benefit is less than B, then it is felt not to be worth it). A narrow prediction interval that lies completely on the beneficial side of a clinically relevant threshold increases confidence in an intervention. A broad prediction interval may indicate the existence of settings where the treatment has a suboptimal and possibly even harmful effect. In more than 70% of statistically significant meta-analyses of the Cochrane Database with some estimated or assumed between-study heterogeneity the prediction intervals crossed the no-effect threshold, indicating that there are settings where those treatments will have no effect or even an effect in the opposite direction. In 20.3% of those meta-analyses the prediction interval even contained the opposite effect of the summary estimate, for example an OR of 0.5 when the summary point estimate was an OR of 2. This occurred most frequently for meta-analyses with a continuous outcome, probably because heterogeneity can be more prominent in many topics where outcomes are assessed on continuous scales; higher heterogeneity for the continuous outcomes was also observed in the full set of 3263 meta-analyses.[13] It was also slightly more common for meta-analyses based on more than six studies, probably because such meta-analyses have more power to detect smaller effects, which means that also the opposite effects will be smaller.

Graham en Moran[19] evaluated prediction intervals in 72 meta-analyses with a dichotomous outcome in critical care published between 2002 and 2010. They found a higher percentage of significant meta-analyses (50/72, 69.4%), compared to 28.5% (572/2009) in our set of meta-analyses with an odds ratio outcome. The difference may be caused by publication bias, the higher number of primary studies in their sample (medium 9 versus 4 in our set[13]), and by their use of the DerSimonian-Laird approach which can result in too many statistically significant findings, whereas we used the HKSJ approach.[14] However, results with respect to the prediction interval were remarkably similar. In 32 (64.0%) of their 50 significant meta-analyses the 95% prediction interval included the null, similar to 65.8% in our dataset. Seven (14.0%) of their 50 meta-analyses suggested a high probability of exact reversal of the efficacy or harm, similar to 12.3% of our meta-analyses where the prediction interval contained the opposite effect, despite the fact that they used a different definitions for possible "harm" and that they did not mention whether there was positive between-study heterogeneity in their significant meta-analyses.

It is straightforward to calculate a prediction interval if we can assume that the effects are normally distributed and that τ^2 is known and stable across studies. However, one should realize that the prediction interval is dependent on this assumption and on the precisions of the estimated τ^2 and study effect, and will be imprecise if the number of studies in the meta-analysis is small. If the number of studies is large, estimates will be more precise and the normality of the distribution of τ^2 can be empirically evaluated. A final caveat is that the uncertainty conveyed by the prediction interval pertains to the uncertainty about the extent to which future studies are similar (exchangeable) to those that have already been done, but this applies to all inferences from a meta-analysis. If the future studies evaluate patients and settings that are entirely different from what was evaluated in past studies, this

exchangeability is questionable and uncertainty may be even more prominent than what the prediction interval conveys. In practical terms, if the patients treated by a physician are considered to be very different from the patients seen in all studies that have been done in the past, even the prediction interval cannot tell us what we might expect for these patients.

POWER CALCULATIONS FOR A FUTURE STUDY

Meta-analysis results can also be used for power calculations for a new study. However, the expected true effect in a new study is not necessarily equal to the point estimate of the meta-analysis: it can be any of the values in the prediction interval. In case of heterogeneity an apparent power of 80% based on the point estimate will be overly optimistic because the power function is asymmetric. If the true study effect is larger than the point estimate the real power of the study will be higher, up to a maximum of 100%, but if the effect is smaller the power may decrease substantially, even to 5% or less in case of a null effect. Consequently the expected power of a new study in case of heterogeneity will be lower than 80% (formula 4 appendix). For example, if the prediction interval shows that 30% of future studies may have a true null or negative effect, the power can never be much larger than 70%. The sample size should be increased to compensate for this loss in power, see also Roloff et al.[20]

Summarizing, the prediction interval reflects the variation in true treatment effects over different settings, including what effect is to be expected in future patients such as the patients that a clinician is interested to treat. Therefore it should be routinely reported in addition to the summary effect and its confidence interval, and used as a main tool for interpreting evidence, to enable more informed clinical decision making.

APPENDIX

Formula 1 Prediction interval

In order to calculate the 95% prediction interval, the summary meta-analysis estimate M, the two sided critical t-value $t_{0.05/2,\,k-1}$ and the standard deviation for the prediction interval SD_{PI} are needed. Here, *t* is the two-sided critical t-value that can be calculated via http://www.danielsoper.com/statcalc3/calc.aspx?id=10. Fill in DF=k-1 and probability level 0.025, with *k* the number of studies in the meta-analysis. SD_{PI} is the standard deviation of the prediction interval: $SD_{PI} = \sqrt{(\tau^2+SE^2)}$,where τ^2 is the estimated heterogeneity and SE is the standard error of M[1 16]. If the SE was not reported, it can be approximated by dividing the distance between the limits of the 95% CI of the SMD by 3.92. The lower and upper limits of the 95% prediction interval are equal to $M \pm t_{0.05/2,\,k-1} \times SD_{PI}$. Of course it is possible to estimate prediction intervals with a different coverage, e.g. an 80% prediction interval would be based on $t_{0.20/2,6}$.

Estimations for ORs, risk ratios and hazard ratios are generally performed on the natural logarithm scale. As an example we take the calculation of a 95% prediction interval for an OR of 2.28 with a 95% CI from 1.05 to 4.96, $\tau^2 = 0.353$ and $k=7$. The prediction interval will first be estimated on log scale. Note that the reported τ^2 is in general already the heterogeneity for log OR, not for OR, and can thus be used directly in the calculations. The SE of the log OR is calculated by dividing the distance between the log of the limits of the 95% CI of the OR by 3.92. This results in SE=0.318.

The lower and upper limits of the 95% prediction interval for the log OR are $\log(2.28) \pm 2.45\sqrt{(0.353 + 0.318^2)}$. The value 2.45 results from the $t_{0.05/2}$-distribution with 6 DF. Finally, we exponentiate the limits to return to the OR scale. The resulting prediction interval ranges

from 0.44 to 11.86, and can be interpreted as the 95% range of true ORs to be expected in similar studies.

Formula 2 Probability that effect is larger than threshold D

The probability P that the true effect in a new study will be below a threshold D (e.g. the null effect) can be calculated with the left-tail cumulative t-distribution with k-1 degrees of freedom. The probability that the effect is above D equals 1 – P.

In our example on nasal polyps the probability that the SMD ≥ 0 can be estimated as follows:

1. Start to calculate the probability P that a true SMD ≤ 0. This is equivalent to the probability that a t-value ≤ T, where T is equal to $(D - M)/SD_{PI}$, with summary treatment effect M= -0.51, SD_{PI} = 0.425 and D=0. This results in T = 1.207, with 6 degrees of freedom (DF).
2. The probability P can be calculated online at http://www.danielsoper.com/statcalc3/calc.aspx?id=41. Fill in t value = 1.207 and DF = 6. The one-tailed probability P(t ≤1.207)= 0.864.
3. We want the probability that the SMD ≥0, this is 1 – P= 0.136.

In the example on the OR (see formula 1), if we are interested in the probability of a null or negative effect, we are interested in the probability that a true OR ≤ 1. For ORs, calculations must be based on the ln OR, with M= ln(2.28)=0.824, SD_{PI} = 0.674, and DF=6. A true OR ≤ 1 corresponds to a true ln OR ≤ 0. Fill in T = (0-0.824)/0.674 = -1.223 and DF=6. The probability that a true OR ≤ 1 is equal to 0.134.

Formula 3 Prediction interval starting with I^2

In order to calculate prediction intervals starting with an assumed I^2 value (as percentage), we first calculated the corresponding τ^2value:

$$\tau^2 = s^2 \frac{I^2}{100 - I^2}$$

with s^2 the typical study variance, equal to $\frac{\sum w_i (k-1)}{(\sum W_i)^2 - \sum w_i^2}$, and w_i equal to the inverse of the study variance of study i (i=1..k) and k the number of studies.[21] Subsequently formula 1 can be applied.

Formula 4 Power of a future study

Usually sample size calculations are performed without consideration of the heterogeneity. If we do take into account the heterogeneity, the expected power, i.e. the probability that a new study with N patients will have a positive result at significance level α, given values for the standard error *s* of the new study and μ and τ^2 as above, can be approximated with the delta method if τ^2 is not too large:

$$E(power) = g(\mu) + 0.5\,\tau^2\,g''(\mu)$$

where g is the power at the meta-analysis summary estimate μ, and $g''(\mu)$ is the second derivative of g at μ. For $g''(\mu)$ we can take the second derivative of the normal cumulative distribution function if N is sufficiently large.

This results in $g''(\mu) = \frac{z_\mu e^{-0.5 z_\mu^2}}{s^2\sqrt{2\pi}}$, with $z_\mu = \frac{1.96s - \mu}{s}$.

If the sample size N of the new study is such that the power for an effect of size μ is 80%, the expected power of the study will be smaller than 80% if τ^2 is positive, because the corresponding value of z_μ is negative.

TABLES AND FIGURES

Table 1 Some frequently used measures for heterogeneity

Table 2 Proportion of statistically significant meta-analyses where both the 95% confidence and prediction intervals excluded the null

Figure 1 Forest plot of the standardized mean difference in symptom scores in nasal polyps. Steroids versus placebo, Analysis 1.1 in Cochrane Review CD006549.[10]

Note that our results differ from the original analysis, as we used a random-effects analysis with the Hartung-Knapp/Sidik-Jonkman adjustment[14] and the empirical Bayes estimator for τ^2.

WEB TABLES

Table W1 Proportion of statistically significant meta-analyses where both the 95% confidence and prediction intervals excluded the null

Separately for dichotomous and continuous outcomes and 2-6 vs. >6 studies

Table 1

Measure	Advantages	Disadvantages
τ^2	• τ (the square root of τ^2) is the standard deviation of the between-study variation on the scale of the original outcome • τ^2 is the direct estimate of the between-study variation and therefore useful in calculations, e.g. for the prediction interval	• A direct clinical interpretation based on τ^2 is difficult, especially when τ^2 belongs to outcomes that were analyzed on log-scale, e.g. odds ratios • When the τ^2 estimate is based on only a few studies it will be imprecise
I^2	• I^2 presents the inconsistency between the study results and quantifies the proportion of observed dispersion that is real, i.e. due to between-study differences and not due to random error[2 3] • I^2 reflects the extent of overlap of the confidence intervals of the study-effects • I^2 represents the inconsistency always on a scale between 0 and 100, therefore it can be compared with suggested limits for low or high inconsistency[11]	• A direct clinical interpretation of I^2 is difficult. • I^2 is also ambiguous because its size depends on sample size: o with very large studies, even tiny between-study differences in effect size may result in a high I^2 o with small (imprecise) studies, very different treatment effects can yield an I^2 of 0
Confidence interval (CI)	• The CI in a random effects model contains highly probable values for the summary (mean) treatment effect	• The CI gives no information on the range of true treatment effects that are likely to be seen in other settings, e.g. in the next study or in the patients a clinician wants to treat in her clinic

Prediction interval	• The prediction interval in a random effects model contains highly probable values for the true treatment effects in future settings, if those settings are similar to the settings in the meta-analysis • The values in the interval can be compared with clinically relevant thresholds to see whether they correspond to benefit, null effects or harm • The prediction interval can be used to estimate the probability that the treatment in a future setting will have a true positive or negative effect, and to perform better power calculations	• Conclusions drawn from the prediction interval are based on the assumption that τ^2 and the study effects are normally distributed • The estimate of the prediction interval will be imprecise if the estimates of the summary effect and the τ^2 are imprecise, for example if they are based on only a few studies and if these studies are small

Table 2

Statistically significant meta-analyses	Estimated heterogeneity I^2				
	$I^2=0$ [a)]	$I^2>0$	>0-30%	30-60%	>60%
N	441	479	123	150	206
Both 95% CI and 95% PI excluded null (n (%))	112 (25.4)	132 (27.6)	88 (71.5)	39 (26.0)	5 (2.4)

CI: confidence interval; PI: prediction interval. [a)] When the estimated heterogeneity I^2 was equal to 0, I^2=20% was imputed for the calculation of the prediction interval.

Figure 1

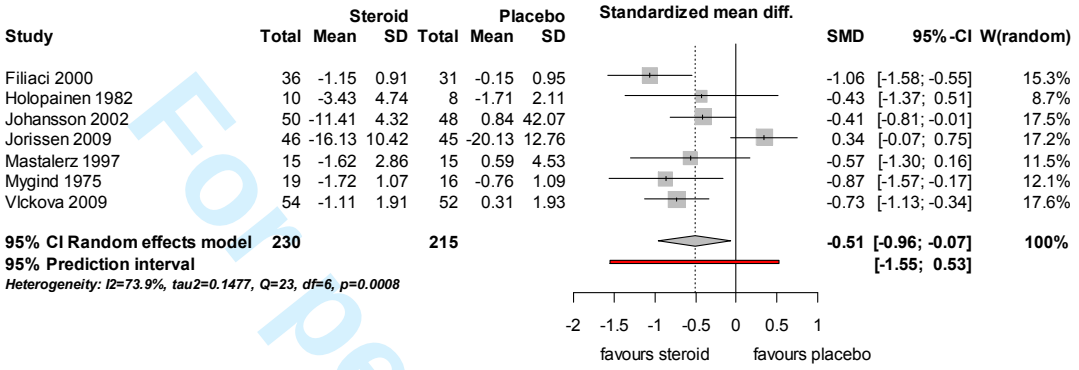

Table W1

	MAs with 2- 6 studies				MAs with >6 studies				all MAs				
	I^2=0	I^2<30	30-60	>60	I^2=0	<30	30-60	>60	I^2=0	<30	30-60	>60	I^2>0
All meta-analyses (N=3263)													
MA stat. significant (N)	322	44	59	59	119	79	91	147	441	123	150	206	479
Both 95% CI and 95% PI excluded the null [a)] (N (%))	74 (23.0)	32 (77.7)	17 (18.8)	1 (1.7)	38 (31.9)	56 (70.9)	22 (24.2)	4 (2.7)	112 (25.4)	88 (71.5)	39 (26.0)	5 (2.4)	132 (27.6)
MAs with dichotomous outcome (N=2009)													
MA stat. significant (N)	210	32	30	20	102	56	66	56	312	88	96	76	260
Both 95% CI and 95% PI excluded the null [a)] (N (%))	50 (23.8)	24 (75.0)	7 (23.3)	1 (5.0)	29 (28.4)	37 (66.1)	16 (24.2)	4 (7.1)	79 (25.3)	61 (69.3)	23 (24.0)	5 (6.6)	89 (34.2)
MAs with continuous outcome (N=1254)													
MA stat. significant (N)	112	12	29	39	17	23	25	91	129	35	54	130	219
Both 95% CI and 95% PI excluded the null [a)] (N (%))	24 (21.4)	8 (66.7)	10 (34.5)	0 (0.0)	9 (52.9)	19 (82.6)	6 (24.0)	0 (0.0)	33 (25.6)	27 (77.1)	16 (29.6)	0 (0.0)	43 (19.6)

MA: meta-analysis; CI= 95% confidence interval; PI= 95% prediction interval; [a)] When the estimated heterogeneity I^2 was equal to 0, I^2=20% was imputed for the calculation of the prediction interval.

COMPETING INTERESTS

No, there are no competing interests

CONTRIBUTORSHIP STATEMENT

JIH originated the idea for this study together with JJG. JIH drafted the manuscript and conducted the data analysis. All authors read and critically revised the manuscript for important intellectual content and approved the final manuscript.

DATA SHARING STATEMENT

Only published data from the Cochrane Database Of Systematic Reviews Issues 2009-2013 were used.These were provided by the UK Cochrane Editorial Unit.

References

1. Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. BMJ 2011;**342**.

2. Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? Statistics in Medicine 2002;**21**(11):1559-73.

3. Higgins J, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. BMJ 2003;**327**(7414):557.

4. Higgins J, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. Journal of the Royal Statistical Society: Series A (Statistics in Society) 2009;**172**(1):137-59.

5. Saha S, Chant D, McGrath J. Meta-analyses of the incidence and prevalence of schizophrenia: conceptual and methodological issues. International Journal of Methods in Psychiatric Research 2008;**17**(1):55-61.

6. Borenstein M, Hedges LV, Higgins JPT, et al. *Introduction to Meta-Analysis*. Chichester, UK: Wiley, 2009.

7. Melsen WG, Bootsma MCJ, Rovers MM, et al. The effects of clinical and statistical heterogeneity on the predictive values of results from meta-analyses. Clinical Microbiology and Infection 2014;**20**(2):123-29.

8. Moonesinghe R, Khoury MJ, Liu T, et al. Required sample size and nonreplicability thresholds for heterogeneous genetic associations. Proceedings of the National Academy of Sciences 2008;**105**(2):617-22.

9. Chiolero A, Santschi V, Burnand B, et al. Meta-analyses: with confidence or prediction intervals? European Journal of Epidemiology 2012;**27**(10):823-25.

10. Kalish L, Snidvongs K, Sivasubramaniam R, et al. Topical steroids for nasal polyps. Cochrane Database Syst Rev 2012;**12**:CD006549.

11. Higgins JPT, Green S, Collaboration C. *Cochrane handbook for systematic reviews of interventions*: Wiley Online Library, 2008.

12. Guddat C, Grouven U, Bender R, et al. A note on the graphical presentation of prediction intervals in random-effects meta-analyses. Systematic Reviews 2012;**1**(1):34.

13. IntHout J, Ioannidis JPA, Borm GF, et al. Small studies are more heterogeneous than large ones: a meta-meta-analysis. Journal of Clinical Epidemiology 2015;**68**(8):860-69.

14. IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC medical research methodology 2014;**14**(1):25.

15. R: A language and environment for statistical computing. Retrieved from http://www.R-project.org/. [program]. Vienna, Austria: R Foundation for Statistical Computing, 2014.

16. Viechtbauer W. Conducting meta-analyses in R with the metafor package. Journal of Statistical Software 2010;**36**(3):1-48.

17. meta: General Package for Meta-Analysis. R package version 4.1-0. http://CRAN.R-project.org/package=meta [program], 2015.

18. Ioannidis JPA, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analyses. BMJ 2007;**335**(7626):914-16.

19. Graham PL, Moran JL. Robust meta-analytic conclusions mandate the provision of prediction intervals in meta-analysis summaries. Journal of Clinical Epidemiology 2012;**65**(5):503-10.

20. Roloff V, Higgins J, Sutton AJ. Planning future studies based on the conditional power of a meta-analysis. Statistics in Medicine 2013;**32**(1):11-24.

21. Bowden J, Tierney JF, Copas AJ, et al. Quantifying, displaying and accounting for heterogeneity in the meta-analysis of RCTs using standard and generalised Q statistics. BMC Medical Research Methodology 2011;**11**(1):41.

BMJ Open

A plea for routinely presenting prediction intervals in meta-analysis

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2015-010247.R1
Article Type:	Research
Date Submitted by the Author:	07-Mar-2016
Complete List of Authors:	IntHout, Joanna; Radboud university medical center, Radboud Institute for Health Sciences, Health Evidence Ioannidis, John; Stanford University, Stanford Prevention Research Center, Department of Medicine and Department of Health Research and Policy Rovers, Maroeska; Radboud university medical center Goeman, Jelle; Radboud university medical center, Radboud Institute for Health Sciences, Health Evidence
Primary Subject Heading:	Research methods
Secondary Subject Heading:	Communication, Epidemiology, Evidence based practice
Keywords:	Meta-analysis, Prediction interval, Heterogeneity, Random effects, Clinical trial, Cochrane Database of Systematic Reviews

A plea for routinely presenting prediction intervals in meta-analysis

Authors: Joanna IntHout*, John PA Ioannidis, Maroeska M Rovers, Jelle J Goeman

Joanna IntHout[1*]
* Corresponding author
Email: Joanna.IntHout@radboudumc.nl

John PA Ioannidis[2,3,4,5]
Email: JIoannid@stanford.edu

Maroeska M Rovers[1]
Email: Maroeska.Rovers@radboudumc.nl

Jelle J Goeman[1]
Email: Jelle.Goeman@radboudumc.nl

[1] Radboud university medical center, Radboud Institute for Health Sciences (RIHS), Mailbox 133, P.O. box 9101, 6500 HB Nijmegen, The Netherlands

[2] Stanford Prevention Research Center, Department of Medicine, Stanford University School of Humanities and Sciences, Stanford, CA 94305, USA

[3] Department of Health Research and Policy, Stanford University School of Medicine, Stanford, CA 94305, USA

[4] Department of Statistics, Stanford University School of Humanities and Sciences, Stanford, CA 94305, USA

[5] Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA 94305, USA

ABSTRACT

Objectives: Evaluating the variation in the strength of the effect across studies is a key feature of meta-analyses. This variability is reflected by measures like τ^2 or I^2 but their clinical interpretation is not straightforward. A prediction interval is less complicated: it presents the expected range of true effects in similar studies. We aimed to show the advantages of having the prediction interval routinely reported in meta-analyses.

Design: We show how the prediction interval can help understand the uncertainty about whether an intervention works or not. To evaluate the implications of using this interval to interpret the results, we selected the first meta-analysis per intervention review of the Cochrane Database of Systematic Reviews Issues 2009-2013 with a dichotomous (n=2009) or continuous (n=1254) outcome, and generated 95% prediction intervals for them.

Results: In 72.4% of 479 statistically significant (random effects $p<0.05$) meta-analyses in the Cochrane Database 2009-2013 with heterogeneity ($I^2>0$), the 95% prediction interval suggested that the intervention effect could be null or even be in the opposite direction. In 20.3% of those 479 meta-analyses, the prediction interval showed that the effect could be completely opposite to the point estimate of the meta-analysis. We demonstrate also how the prediction interval can be used to calculate the probability that a new trial will show a negative effect and to improve the calculations of the power of a new trial.

Conclusions: The prediction interval reflects the variation in treatment effects over different settings, including what effect is to be expected in future patients such as the patients that a clinician is interested to treat. Prediction intervals should be routinely reported to allow more informative inferences in meta-analyses.

Word count: 274

KEYWORDS

Meta-analysis; Prediction interval; Heterogeneity; Random effects; Clinical trial; Cochrane Database of Systematic Reviews;

STRENGTHS AND LIMITATIONS OF THIS STUDY

- In many meta-analyses there is large variation in the strength of the effect.
- The prediction interval helps in the clinical interpretation of the heterogeneity by estimating what true treatment effects can be expected in future settings.
- In case of heterogeneity, prediction intervals will show a wider range of expected treatment effects than confidence intervals, and thus may lead to different conclusions. This occurred in over 70% of statistically significant meta-analyses with heterogeneity of the Cochrane Database of Systematic Reviews. Completely opposite effects were not excluded in over 20% of those meta-analyses.
- Prediction intervals should be routinely reported to allow more informative inferences in meta-analyses.
- Limitations are that the calculations and inferences for the prediction interval are based on the normality assumption, which is difficult to ensure. Further, the interval will be imprecise if the estimates of the summary effect and the between-study heterogeneity are imprecise, for example if they are based on only a few, small studies. Inferences based on the prediction interval are only valid for settings that are similar (exchangeable) to those on which the meta-analysis is based.

MAIN PAPER

INTRODUCTION

Interventions may have heterogeneous effects across studies because of differences in study populations, interventions, follow-up length, or other factors like publication bias.[1] Nevertheless, the usual reporting of a meta-analysis is focused on the summary effect size combined with a confidence interval (CI) and p-value. Typically also some measure of the between-study heterogeneity is presented such as τ^2 or the inconsistency measure I^2.[2,3] However, neither of these two metrics can readily point to the clinical implications of the observed heterogeneity. Our objective in the current article is to show the potential advantages of obtaining and reporting the prediction interval routinely in meta-analyses because its clinical meaning is much more straightforward. The prediction interval presents the heterogeneity in the same metric as the original effect size measure, in contrast to τ^2 or I^2. Reporting a prediction interval in addition to the summary estimate and confidence interval will illustrate which range of true effects can be expected in future settings. We describe its merits and provide working examples to show how it can be calculated.

METHODS

1. INTERPRETATION OF HETEROGENEITY

Between-study variation in the magnitude of treatment effects cannot be neglected. One of the main merits of a meta-analysis may even be that it reveals the variation of effects in different studies.[4] Therefore summarizing the findings of a meta-analysis in a single summary value sacrifices potentially informative variation.[5] However, the information that can be directly retrieved from τ^2 and I^2 with respect to the variation in the effects is limited. The clinical interpretation of I^2 is ambiguous: a high I^2 does not necessarily imply that the study effects are dispersed over a wide range[6] and a low I^2 might correspond to high dispersion[7], because I^2

depends on sample size of the included studies[8]. With very large (highly precise) studies, even tiny differences in effect size may result in a high I^2, while with small (imprecise) studies, very different treatment effects can yield an I^2 of 0. Dispersion in treatment effects is better reflected by τ because τ is the standard deviation of the between-study effects. One could for example estimate the ratio of the effect size over τ, which can convey how many times larger the treatment effect is compared to the standard deviation of the effect across studies.[9] But this may still be not very intuitive to a clinical reader. Another popular way to express variation in effect sizes is the CI, e.g. the 95% CI. The CI in a random effects model contains highly probable values for the summary treatment effect. However, it does not convey what range of treatment effects are likely to be seen in other patients, e.g. in the next study or in the patients a clinician wants to treat in her clinic.

2. PREDICTION INTERVALS

Not so often reported but much more insightful is the prediction interval.[10] A prediction interval always presents the heterogeneity on the same scale as the original outcomes, in contrast to τ (e.g. in case of odds ratios), τ^2 or I^2. A 95% prediction interval estimates where the true effects are to be expected for 95% of similar (exchangeable) studies that might be conducted in the future.[4] Therefore it is well suited to evaluate the variability of the effect of an intervention over different settings. For example, in a meta-analysis on sedentary time in adults and the association with diabetes, cardiovascular disease and death, confidence intervals were thought to represent insufficiently the different study populations. Therefore also prediction intervals were reported.[11] In the absence of between-study heterogeneity, the prediction interval coincides with the respective CI. However, in case of heterogeneity a prediction interval covers a wider range than a CI. Consequently, in case of a statistically significant effect (where all values of the 95% CI are on the same side of the null) the

corresponding 95% prediction interval may indicate that values are possible on both sides of the null. This means that there will be settings where conclusions based on CIs will not hold. In the same framework, one can also calculate the probability that the true effect will be harmful (on the other side of the null) in a next study. Table 1 presents an overview of measures of between-study heterogeneity.

Table 1 approximately here

3. EXAMPLE: TOPICAL STEROIDS FOR NASAL POLYPS

A 2012 review on the use of topical steroids for treatment of chronic rhinosinusitis with nasal polyps, based on seven randomized studies, resulted in a larger decrease in overall symptom scores in favor of steroids compared to placebo[12]. This is reflected by a standardized mean difference (SMD) of -0.51, with a 95% CI from -0.96 to -0.07 (Figure 1). The I^2 is 73.9% (95% CI, 44.2% to 87.8%), which can be considered substantial heterogeneity[13], and the estimated τ^2 is 0.148. Notwithstanding these numbers, it is difficult to evaluate what the clinical consequences of this heterogeneity may be for future settings.

In order to estimate the prediction interval for the SMD we need the point estimate of the SMD, its standard error (SE) and the estimated τ^2. We derive the SE from the 95% CI of the SMD (formula 1 appendix), which results in an SE of 0.227. We can calculate the standard deviation of the prediction interval SD_{PI} as $\sqrt{(0.148 + 0.227^2)}$ and the lower and upper limit of the 95% prediction interval as $-0.51 \pm 2.45 \times SD_{PI}$. The value 2.45 results from the $t_{1\text{-}0.05/2,6}$ distribution. Prediction intervals with a different coverage could be calculated by using a different t-value, e.g. $t_{1\text{-}0.20/2,6}$ for an 80% prediction interval (formula 1 appendix).

The resulting prediction interval, ranging from -1.60 to 0.58, can be interpreted as the 95% range of true SMDs to be expected in similar studies. We present it in Figure 1 as a rectangle below the diamond for the 95% CI.[14] The prediction interval contains values below zero, which correspond to a decrease in symptom scores of at best approximately 1.6 SD after steroid use compared to placebo. But it also contains values above zero which means that the steroids may exhibit no or even a harmful effect (SMD>0) in some settings, with a (95%) worst case increase in SMD of 0.58. Consequently, the effect in a new study may be even the exact opposite to the summary point estimate of the meta-analysis, i.e. an increase of 0.51 instead of a decrease of -0.51 may occur. The estimated probability that the true effect of the steroids will be null or higher in a new study is equal to 14.7%, based on the t-distribution with 6 degrees of freedom (formula 2 appendix).

Figure 1 approximately here

4. COCHRANE DATABASE

In order to investigate how often there is a discrepancy in conclusions based on prediction intervals and CIs we evaluated this in statistically significant meta-analyses ($p<0.05$ by random effects calculations) of the Cochrane Database of Systematic Reviews Issues 2009-2013, kindly provided by the UK Cochrane Editorial Unit. To avoid subjectivity in the selection we used the first meta-analysis with a dichotomous or continuous outcome and based on at least two studies in the Data and Analyses section when these studies were also combined in the original review, as we wanted to reflect the status quo as precise as possible. Details can be found in another paper.[15] In brief, of a total of 3263 meta-analyses, 920 were statistically significant: 479 with an estimated $I^2>0$ and 441 with an estimated $I^2=0$.

5. CALCULATIONS

We used the Hartung-Knapp/Sidik-Jonkman[16] (HKSJ) random effects meta-analysis approach combined with the empirical Bayes estimator for τ^2. We estimated τ^2 for all meta-analyses, even when the authors originally performed a fixed effects analysis. Prediction intervals were calculated according to formula 1 (Appendix). We categorized the statistically significant meta-analyses with heterogeneity ($\tau^2>0$) by number of studies (2-6 studies or >6) and heterogeneity ($I^2<30\%$, 30-60%, or >60%, based on the Cochrane Handbook[13] stating that an I^2 between 30% and 60% corresponds to moderate heterogeneity). For significant meta-analyses where the heterogeneity estimate was zero, we assessed the impact of possibly low but non-zero heterogeneity by assuming an I^2 of 20%, calculating prediction intervals using formula 3 (appendix). Categorical outcomes were compared between groups by means of the chi-square test. We used R software[17] version 3.1.2 and the R packages metafor[18] version 1.9-5 and meta[19] version 4.1-0.

RESULTS

Overall, 132 (27.6%) of the 479 statistically significant meta-analyses with an $I^2 > 0$ had both the 95% CI and the 95% prediction interval excluding the null effect (Table 2). Consequently, almost three-quarter (347, 72.4%) had a prediction interval that contained the null effect. This means that it is likely that for these comparisons some patient populations might experience null effects or effects in the opposite direction, i.e. a treatment might be more harmful than the comparator even though the point estimate suggests benefit (or vice versa). Not surprisingly, significant meta-analyses with low heterogeneity more often had prediction intervals that excluded the null than meta-analyses with high heterogeneity. The percentage of prediction intervals containing the null effect was slightly higher for meta-analyses with a continuous outcome (80.4%) than for those with a dichotomous outcome (65.8%) ($p<0.001$), but not

significantly different for meta-analyses based on more than six studies (74.1%) than for those with at most six studies (69.1%) (p=0.25). (Web Table W1).

Of the 347 meta-analyses with a prediction interval that contained the null or opposite effect, 199 (57.3%) had also at least one study with an opposite effect. This happened more often in meta-analyses with more than six studies (181/235, 77.0%) than in those based on at most six studies (18/102, 17.6%). Especially in meta-analyses with few studies and substantial heterogeneity, the prediction interval was wider than the range of study outcomes. The opposite (i.e. a smaller prediction interval) occurred in meta-analyses based on many studies and with low estimated heterogeneity. Results for meta-analyses with dichotomous and continuous outcomes were not notably different.

PREDICTION INTERVALS CONTAINING THE OPPOSITE EFFECT

If the prediction interval just includes the null effect, this may be less worrying than when it contains the exact opposite effect of the pooled summary effect, e.g. if it contains an OR of 0.5 when the meta-analysis summary estimate is an OR of 2, or if it contains an SMD of -0.7 when the summary estimate was 0.7. Of the 479 significant meta-analyses with an $I^2 > 0$, 97 (20.3%) had a prediction interval that contained the opposite effect. This percentage was higher for the meta-analyses with a continuous outcome (65/219, 29.7%) than for those with a dichotomous outcome (32/260, 12.3%) (p<0.001). It occurred also more frequently in meta-analyses with more than 6 primary studies (57/139, 41.0% and 30/178, 20.3% for meta-analyses with a continuous or dichotomous outcome, respectively) than for those based on at most 6 studies (8/80, 10.0% and 2/82, 2.4%) (p<0.001 and p=0.001, respectively).

Table 2 approximately here

META-ANALYSES WITH ESTIMATED I^2=0

A substantial part of meta-analyses have an estimated I^2 of 0. However, there is typically very large uncertainty about the exact amount of heterogeneity and this is demonstrated by very large 95% CIs for the values of I^2.[20] The same applies to τ: an estimate of 0 is often accompanied by large uncertainty. The true I^2 and τ are unlikely to ever be exactly 0, although low values are possible. To assess the impact of possibly low but non-zero heterogeneity among the 441 Cochrane meta-analyses with estimated I^2=0 and statistically significant results, we imputed an I^2=20% (suggestive of low between-study heterogeneity). Under this assumption, in 329 (74.6%) of these 441 meta-analyses the 95% prediction interval would span both sides of the null (Table 2), similar for meta-analyses with a dichotomous (74.7%) or continuous (74.4%) outcome (Web Table W1). This is a sensitivity analysis that is useful to perform to see whether the inferences of a meta-analysis that seemingly does not have detectable heterogeneity may be influenced by even a small amount of heterogeneity.

DISCUSSION AND OUTLOOK

In meta-analyses a CI is inadequate for clinical decision making because it only summarizes the average effect for the average study. The prediction interval is more informative as it shows the range of possible effects in relation to harm and clinical benefit thresholds. While we have focused on the situation where the separating threshold is the null, a different threshold may be considered. For example, in the prediction interval framework one can calculate the probability that an effect is larger than B, where B may be a clinically meaningful effect (if the treatment benefit is less than B, then it is felt not to be worth it). A narrow prediction interval that lies completely on the beneficial side of a clinically relevant threshold increases confidence in an intervention. A broad prediction interval may indicate the

existence of settings where the treatment has a suboptimal and possibly even harmful effect. In more than 70% of statistically significant meta-analyses of the Cochrane Database with some estimated or assumed between-study heterogeneity the prediction intervals crossed the no-effect threshold, indicating that there are settings where those treatments will have no effect or even an effect in the opposite direction. In 20.3% of those meta-analyses the prediction interval even contained the opposite effect of the summary estimate, for example an OR of 0.5 when the summary point estimate was an OR of 2. This occurred most frequently for meta-analyses with a continuous outcome, probably because heterogeneity can be more prominent in many topics where outcomes are assessed on continuous scales; higher heterogeneity for the continuous outcomes was also observed in the full set of 3263 meta-analyses.[15] It was also slightly more common for meta-analyses based on more than six studies, probably because such meta-analyses have more power to detect smaller effects, which means that also the opposite effects will be smaller.

Graham en Moran[21] evaluated prediction intervals in 72 meta-analyses with a dichotomous outcome in critical care published between 2002 and 2010. They found a higher percentage of significant meta-analyses (50/72, 69.4%), compared to 28.5% (572/2009) in our set of meta-analyses with an odds ratio outcome. The difference may be caused by publication bias, the higher number of primary studies in their sample (median 9 versus 4 in our set[15]), and by their use of the DerSimonian-Laird approach which can result in too many statistically significant findings, whereas we used the HKSJ approach.[16] However, results with respect to the prediction interval were remarkably similar. In 32 (64.0%) of their 50 significant meta-analyses the 95% prediction interval included the null, similar to 65.8% in our dataset. Seven (14.0%) of their 50 meta-analyses suggested a high probability of exact reversal of the efficacy or harm, similar to 12.3% of our meta-analyses where the prediction interval

contained the opposite effect, despite the fact that they used a different definition for possible "harm" and that they did not mention whether there was positive between-study heterogeneity in their significant meta-analyses.

It is straightforward to calculate a prediction interval if we can assume that the effects are normally distributed and that τ^2 is known and stable across studies. However, one should realize that the prediction interval is dependent on this assumption and on the precisions of the estimated τ^2 and study effect, and will be imprecise if the number of studies in the meta-analysis is small. If the number of studies is large, estimates will be more precise and the normality of the distribution of τ^2 can be empirically evaluated. A final caveat is that the uncertainty conveyed by the prediction interval pertains to the uncertainty about the extent to which future studies are similar (exchangeable) to those that have already been done, but this applies to all inferences from a meta-analysis. If the future studies evaluate patients and settings that are entirely different from what was evaluated in past studies, this exchangeability is questionable and uncertainty may be even more prominent than what the prediction interval conveys. In practical terms, if the patients treated by a physician are considered to be very different from the patients seen in all studies that have been done in the past, even the prediction interval cannot tell us what we might expect for these patients.

POWER CALCULATIONS FOR A FUTURE STUDY

Meta-analysis results can also be used for power calculations for a new study. However, the expected true effect in a new study is not necessarily equal to the point estimate of the meta-analysis: it can be any of the values in the prediction interval. In case of heterogeneity the probability of a statistically significant result in a new study may differ substantially from an apparent power of 80% based on the point estimate. The latter will be overly optimistic

because the power function is asymmetric. If the true study effect is larger than the point estimate the real probability of a significant study will be higher, up to a maximum of 100%, but if the effect is smaller the probability may decrease substantially, even to 5% or less in case of a null effect. Consequently the expected probability of a significant new study in case of heterogeneity will be lower than 80% (formula 4 appendix). For example, if the prediction interval shows that 30% of future studies may have a true null or negative effect, the probability of a significant new study can never be much larger than 70%. The sample size should be increased to compensate for this loss, see also Roloff et al.[22]

Summarizing, the prediction interval reflects the variation in true treatment effects over different settings, including what effect is to be expected in future patients such as the patients that a clinician is interested to treat. Therefore it should be routinely reported in addition to the summary effect and its confidence interval, and used as a main tool for interpreting evidence, to enable more informed clinical decision making.

AUTHORS' CONTRIBUTIONS

JIH originated the idea for this study together with JJG. JIH drafted the manuscript and conducted the data analysis. All authors read and critically revised the manuscript for important intellectual content and approved the final manuscript.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

COMPETING INTERESTS

We have read and understood the BMJ Group policy on declaration of interests and declare the following interests: none.

DATA SHARING STATEMENT

Datasets are available upon request from the corresponding author.

TABLES AND FIGURES

Table 1 Some frequently used measures for heterogeneity

Table 2 Proportion of statistically significant meta-analyses where both the 95% confidence and prediction intervals excluded the null

Figure 1 Forest plot of the standardized mean difference in symptom scores in nasal polyps. Steroids versus placebo, Analysis 1.1 in Cochrane Review CD006549.[12]

Note that our results differ from the original analysis, as we used a random-effects analysis with the Hartung-Knapp/Sidik-Jonkman adjustment[16] and the empirical Bayes estimator for τ^2.

WEB TABLES

Table W1 Proportion of statistically significant meta-analyses where both the 95% confidence and prediction intervals excluded the null

Separately for dichotomous and continuous outcomes and 2-6 vs. >6 studies

Table 1

Measure	Advantages	Disadvantages
τ^2	• τ (the square root of τ^2) is the standard deviation of the between-study variation on the scale of the original outcome • τ^2 is the direct estimate of the between-study variation and therefore useful in calculations, e.g. for the prediction interval	• A direct clinical interpretation based on τ^2 is difficult, especially when τ^2 belongs to outcomes that were analyzed on log-scale, e.g. odds ratios • When the τ^2 estimate is based on only a few studies it will be imprecise
I^2	• I^2 presents the inconsistency between the study results and quantifies the proportion of observed dispersion that is real, i.e. due to between-study differences and not due to random error[2 3] • I^2 reflects the extent of overlap of the confidence intervals of the study-effects • I^2 represents the inconsistency always on a scale between 0 and 100, therefore it can be compared with suggested limits for low or high inconsistency[13]	• A direct clinical interpretation of I^2 is difficult. • I^2 is also ambiguous because its size depends on sample size: o with very large studies, even tiny between-study differences in effect size may result in a high I^2 o with small (imprecise) studies, very different treatment effects can yield an I^2 of 0
Confidence interval (CI)	• The CI in a random effects model contains highly probable values for the summary (mean) treatment effect	• The CI gives no information on the range of true treatment effects that are likely to be seen in other settings, e.g. in the next study or in the patients a clinician wants to treat in her clinic

Prediction interval	• The prediction interval in a random effects model contains highly probable values for the true treatment effects in future settings, if those settings are similar to the settings in the meta-analysis • The values in the interval can be compared with clinically relevant thresholds to see whether they correspond to benefit, null effects or harm • The prediction interval can be used to estimate the probability that the treatment in a future setting will have a true positive or negative effect, and to perform better power calculations	• Conclusions drawn from the prediction interval are based on the assumption that τ^2 and the study effects are normally distributed • The estimate of the prediction interval will be imprecise if the estimates of the summary effect and the τ^2 are imprecise, for example if they are based on only a few studies and if these studies are small

Table 2

Statistically significant meta-analyses	Estimated heterogeneity I^2				
	$I^2=0$ [a)]	$I^2>0$	>0 and <30%	30-60%	>60%
N	441	479	123	150	206
Both 95% CI and 95% PI excluded null, n (%))	112 (25.4)	132 (27.6)	88 (71.5)	39 (26.0)	5 (2.4)

CI: confidence interval; PI: prediction interval. [a)] When the estimated heterogeneity I^2 was equal to 0, I^2=20% was imputed for the calculation of the prediction interval.

References

1. Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. BMJ 2011;**342**.

2. Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? Statistics in Medicine 2002;**21**(11):1559-73.

3. Higgins J, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. BMJ 2003;**327**(7414):557.

4. Higgins J, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. Journal of the Royal Statistical Society: Series A (Statistics in Society) 2009;**172**(1):137-59.

5. Saha S, Chant D, McGrath J. Meta-analyses of the incidence and prevalence of schizophrenia: conceptual and methodological issues. International Journal of Methods in Psychiatric Research 2008;**17**(1):55-61.

6. Borenstein M, Hedges LV, Higgins JPT, et al. *Introduction to Meta-Analysis*. Chichester, UK: Wiley, 2009.

7. Melsen WG, Bootsma MCJ, Rovers MM, et al. The effects of clinical and statistical heterogeneity on the predictive values of results from meta-analyses. Clinical Microbiology and Infection 2014;**20**(2):123-29.

8. Rucker G, Schwarzer G, Carpenter J, et al. Undue reliance on I^2 in assessing heterogeneity may mislead. BMC Medical Research Methodology 2008;**8**(1):79.

9. Moonesinghe R, Khoury MJ, Liu T, et al. Required sample size and nonreplicability thresholds for heterogeneous genetic associations. Proceedings of the National Academy of Sciences 2008;**105**(2):617-22.

10. Chiolero A, Santschi V, Burnand B, et al. Meta-analyses: with confidence or prediction intervals? European Journal of Epidemiology 2012;**27**(10):823-25.

11. Wilmot EG, Edwardson CL, Achana FA, et al. Sedentary time in adults and the association with diabetes, cardiovascular disease and death: systematic review and meta-analysis. Diabetologia 2012;**55**:2895-905.

12. Kalish L, Snidvongs K, Sivasubramaniam R, et al. Topical steroids for nasal polyps. Cochrane Database Syst Rev 2012;**12**:CD006549.

13. Higgins JPT, Green S, Collaboration C. *Cochrane handbook for systematic reviews of interventions*: Wiley Online Library, 2008.

14. Guddat C, Grouven U, Bender R, et al. A note on the graphical presentation of prediction intervals in random-effects meta-analyses. Systematic Reviews 2012;**1**(1):34.

15. IntHout J, Ioannidis JPA, Borm GF, et al. Small studies are more heterogeneous than large ones: a meta-meta-analysis. Journal of Clinical Epidemiology 2015;**68**(8):860-69.

16. IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC medical research methodology 2014;**14**(1):25.

17. R: A language and environment for statistical computing. Retrieved from http://www.R-project.org/. [program]. Vienna, Austria: R Foundation for Statistical Computing, 2014.

18. Viechtbauer W. Conducting meta-analyses in R with the metafor package. Journal of Statistical Software 2010;**36**(3):1-48.

19. meta: General Package for Meta-Analysis. R package version 4.1-0. http://CRAN.R-project.org/package=meta [program], 2015.

20. Ioannidis JPA, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analyses. BMJ 2007;**335**(7626):914-16.

21. Graham PL, Moran JL. Robust meta-analytic conclusions mandate the provision of prediction intervals in meta-analysis summaries. Journal of Clinical Epidemiology 2012;**65**(5):503-10.

22. Roloff V, Higgins J, Sutton AJ. Planning future studies based on the conditional power of a meta-analysis. Statistics in Medicine 2013;**32**(1):11-24.

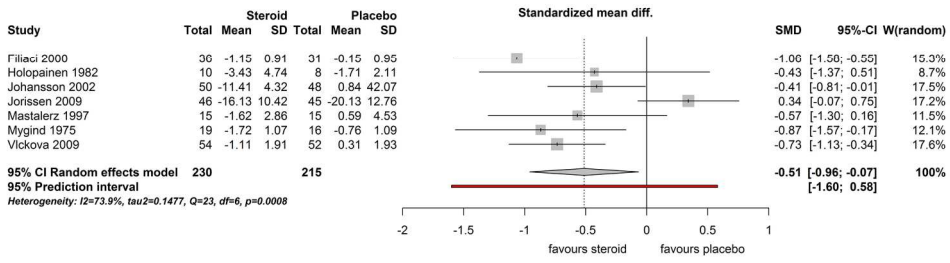

Figure 1 Forest plot of the standardized mean difference in symptom scores in nasal polyps. Steroids versus placebo, Analysis 1.1 in Cochrane Review CD006549.12

Note that our results differ from the original analysis, as we used a random-effects analysis with the Hartung-Knapp/Sidik-Jonkman adjustment and the empirical Bayes estimator for τ^2.

177x88mm (300 x 300 DPI)

APPENDIX

Formula 1 Prediction interval

In order to calculate the 95% prediction interval, the summary meta-analysis estimate M, the two sided critical t-value $t_{1\text{-}0.05/2,\ k\text{-}1}$ and the standard deviation for the prediction interval SD_{PI} are needed. Here, *t* is the two-sided critical t-value that can be calculated via http://www.danielsoper.com/statcalc3/calc.aspx?id=10. Fill in DF=k-1 and probability level 0.025, with *k* the number of studies in the meta-analysis. SD_{PI} is the standard deviation of the prediction interval: $SD_{PI} = \sqrt{(\tau^2+SE^2)}$,where τ^2 is the estimated heterogeneity and SE is the standard error of M[1 18]. If the SE was not reported, it can be approximated by dividing the distance between the limits of the 95% CI of the SMD by 3.92. The lower and upper limits of the 95% prediction interval are equal to $M \pm t_{1\text{-}0.05/2,\ k\text{-}1} \times SD_{PI}$. Of course it is possible to estimate prediction intervals with a different coverage, e.g. an 80% prediction interval would be based on $t_{1\text{-}0.20/2,\ 6}$. Note that the interval is calculated under the assumption that the value of τ^2 is known (and not estimated).

Estimations for ORs, risk ratios and hazard ratios are generally performed on the natural logarithm scale. As an example we take the calculation of a 95% prediction interval for an OR of 2.28 with a 95% CI from 1.05 to 4.96, τ^2= 0.353 and *k*=7. The prediction interval will first be estimated on log scale. Note that the reported τ^2 is in general already the heterogeneity for log OR, not for OR, and can thus be used directly in the calculations. The SE of the log OR is calculated by dividing the distance between the log of the limits of the 95% CI of the OR by 3.92. This results in SE=0.318. The lower and upper limits of the 95% prediction interval for the log OR are $\log(2.28) \pm 2.45\sqrt{(0.353 + 0.318^2)}$. The value 2.45 results from the $t_{1\text{-}0.05/2}$ distribution with 6 DF. Finally, we exponentiate the limits to return to the OR scale. The

resulting prediction interval ranges from 0.44 to 11.86, and can be interpreted as the 95% range of true ORs to be expected in similar studies.

Formula 2 Probability that effect is larger than threshold D

The probability P that the true effect in a new study will be below a threshold D (e.g. the null effect) can be calculated with the left-tail cumulative t-distribution with k-1 degrees of freedom. The probability that the effect is above D equals 1 – P.

In our example on nasal polyps the probability that the SMD ≥ 0 can be estimated as follows:

1. Start to calculate the probability P that a true SMD ≤ 0. This is equivalent to the probability that a t-value ≤ T, where T is equal to $(D - M)/SD_{PI}$, with summary treatment effect M= -0.51, SD_{PI} = 0.425 and D=0. This results in T = 1.207, with 6 degrees of freedom (DF).
2. The probability P can be calculated online at http://www.danielsoper.com/statcalc3/calc.aspx?id=41. Fill in t value = 1.207 and DF = 6. The one-tailed probability P(t ≤1.207)= 0.864.
3. We want the probability that the SMD ≥0, this is 1 – P= 0.136.

In the example on the OR (see formula 1), if we are interested in the probability of a null or negative effect, we are interested in the probability that a true OR ≤ 1. For ORs, calculations must be based on the ln OR, with M= ln(2.28)=0.824, SD_{PI} = 0.674, and DF=6. A true OR ≤ 1 corresponds to a true ln OR ≤ 0. Fill in T = (0-0.824)/0.674 = -1.223 and DF=6. The probability that a true OR ≤ 1 is equal to 0.134.

Formula 3 Prediction interval starting with I^2

In order to calculate prediction intervals starting with an assumed I^2 value (as percentage), we first calculated the corresponding τ^2value:

$$\tau^2 = s^2 \frac{I^2}{100 - I^2}$$

with s^2 the typical study variance, equal to $\frac{\sum w_i (k-1)}{(\sum W_i)^2 - \sum w_i^2}$, and w_i equal to the inverse of the study variance of study i (i=1..k) and k the number of studies.[23] Subsequently formula 1 can be applied.

Formula 4 Power of a future study

Usually sample size calculations are performed without consideration of the heterogeneity. If we do take into account the heterogeneity, the expected power, i.e. the probability that a new study with N patients will have a positive result at significance level α, given values for the standard error *s* of the new study and μ and τ^2 as above, can be approximated with the delta method if τ^2 is not too large:

$$E(power) = g(\mu) + 0.5\,\tau^2\,g''(\mu)$$

where g is the power at the meta-analysis summary estimate μ, and $g''(\mu)$ is the second derivative of g at μ. For $g''(\mu)$ we can take the second derivative of the normal cumulative distribution function if N is sufficiently large.

This results in $g''(\mu) = \frac{z_\mu e^{-0.5 z_\mu^2}}{s^2\sqrt{2\pi}}$, with $z_\mu = \frac{1.96s - \mu}{s}$.

If the sample size N of the new study is such that the power for an effect of size μ is 80%, the expected power of the study will be smaller than 80% if τ^2 is positive, because the corresponding value of z_μ is negative.

WEB TABLE

Table W1 Proportion of statistically significant meta-analyses where both the 95% confidence and prediction intervals excluded the null by estimated I^2

Separately for dichotomous and continuous outcomes and 2-6 vs. >6 studies

From:

IntHout, J., Ioannidis, J., Rovers, M., & Goeman, J. 2016. BMJ Open. A plea for routinely presenting prediction intervals in meta-analysis.

Table W1: Proportion of statistically significant meta-analyses where both the 95% confidence and prediction intervals excluded the null by estimated I^2
Separately for dichotomous and continuous outcomes and 2-6 vs. >6 studies

Estimated heterogeneity I^2 (%)	Meta-analyses with 2-6 studies				Meta-analyses with >6 studies				All meta-analyses				
	0	>0 and <30	30-60	>60	0	>0 and <30	30-60	>60	0	>0 and <30	30-60	>60	I^2>0
All meta-analyses (N=3263)													
MA stat. significant (N)	322	44	59	59	119	79	91	147	441	123	150	206	479
Both 95% CI and 95% PI excluded	74	32	17	1	38	56	22	4	112	88	39	5	132
the null [a)] (N (%))	(23.0)	(77.7)	(18.8)	(1.7)	(31.9)	(70.9)	(24.2)	(2.7)	(25.4)	(71.5)	(26.0)	(2.4)	(27.6)
Meta-analyses with a dichotomous outcome (N=2009)													
MA stat. significant (N)	210	32	30	20	102	56	66	56	312	88	96	76	260
Both 95% CI and 95% PI excluded	50	24	7	1	29	37	16	4	79	61	23	5	89
the null [a)] (N (%))	(23.8)	(75.0)	(23.3)	(5.0)	(28.4)	(66.1)	(24.2)	(7.1)	(25.3)	(69.3)	(24.0)	(6.6)	(34.2)
Meta-analyses with a continuous outcome (N=1254)													
MA stat. significant (N)	112	12	29	39	17	23	25	91	129	35	54	130	219
Both 95% CI and 95% PI excluded	24	8	10	0	9	19	6	0	33	27	16	0	43
the null [a)] (N (%))	(21.4)	(66.7)	(34.5)	(0.0)	(52.9)	(82.6)	(24.0)	(0.0)	(25.6)	(77.1)	(29.6)	(0.0)	(19.6)

MA: meta-analysis; CI= 95% confidence interval; PI= 95% prediction interval;

[a)] When the estimated heterogeneity I^2 was equal to 0, I^2=20% was imputed for the calculation of the prediction interval.