CAN WE TRUST THE RESULTS OF META-ANALYSES? A SYSTEMATIC APPROACH TO SENSITIVITY ANALYSIS IN META-ANALYSES

Rune Elvik
Institute of Transport Economics PO Box 6110 Etterstad N-0602 Oslo, Norway
E-mail: re@toi.no

Abstract

Every meta-analysis involves a number of choices made by the analyst. These choices may refer to, for example, estimator of effect, model for analysis (fixed-effects or random-effects), or the treatment of varying study quality. The choices made can affect the results of the analysis. Every meta-analysis should therefore include a sensitivity analysis, designed to probe how choices made as part of the analysis affects its results. This paper describes a systematic approach to sensitivity analysis in meta-analyses. An index intended to summarise the results of a sensitivity analysis, the robustness score, is developed. The robustness score varies from 0 to 1. A value of 1 indicates that the results of a meta-analysis are very robust, meaning that they are not at all affected by the choices made by the analyst. It is proposed that every meta-analysis should include a sensitivity analysis with respect to: (1) The potential presence of publication bias, (2) Choice of estimator of effect (if relevant), (3) The possible presence of outlier bias (a single result having decisive influence on the summary estimate), (4) Statistical weighting of individual estimates of effect, (5) Assessment of study quality. A recently reported meta-analysis of studies that have evaluated the effects on road safety of daytime running lights for cars is used as a case in order to explain the approach to sensitivity analysis proposed in the paper.

Keywords: Meta-analysis; Sensitivity analysis; Systematic approach; Robustness score
Topic Area: H6 Campbell Collaboration and Effects of Transport and Housing Policy

1. Introduction

As part of any meta-analysis, the analyst has to make a number of choices that may influence the results of the study. These choices include (but are not limited to):

1. Definition of study inclusion criteria, which determine the sample of studies included in the analysis,
2. Choice of summary estimate of effect, which may influence the choice of statistical technique of analysis,
3. Choice of approach to the treatment of heterogeneity in estimates of effect, usually a choice between a fixed-effect and random-effects model,
4. Choices made with respect to the treatment of varying study quality, such as omitting bad studies, scoring studies for quality, etc.

The objective of this paper is to describe a systematic approach to sensitivity analysis as part of meta-analysis. The main research problems the paper seeks to answer are:

1. Which analytic choices made as part of a meta-analysis ought to be included in a sensitivity analysis?

2. How can the effects of analytic choices on the results of a meta-analysis be assessed?

3. How can the results of a sensitivity analysis guide the analytic choices made as part of a meta-analysis?

To help answer these questions, a sensitivity analysis will be performed for a recently reported meta-analysis of studies that have evaluated the effects on road safety of daytime running lights for cars (Elvik, Christensen and Fjeld Olsen 2003). Daytime running lights are headlights used during daytime in order to make cars more conspicuous. The use of daytime running lights is intended to prevent accidents in daytime involving more than one road user.

2. What should be subject to sensitivity analysis?

Analytic choices that can influence the results of a meta-analysis are made both before the analysis starts, and as part of it. Prior to a meta-analysis, a systematic review of relevant studies will be made. As part of this, study inclusion criteria are defined. The definition of these criteria, as well as the effort made in retrieving studies, determines how "complete" a meta-analysis will be with respect to relevant studies. Although an ideal meta-analysis would include every study that has been made about a topic, it is in practice never possible to know for sure that one has succeeded in retrieving all studies. Hence, the first potential source of bias to be tested in a sensitivity analysis is publication bias. Publication bias is found whenever there is an association between study findings and the probability that a study is published. For a review of techniques that have been developed in order to diagnose the presence of publication bias and correcting for it, see Christensen (2003).

In many meta-analyses, there is a choice between several estimators of effect, both within each study and for the summary statistics produced in the meta-analyses. Epidemiological studies, for example, often have a choice to present their findings in terms of the odds ratio, the risk ratio or the risk difference (Deeks 2002). In road safety evaluation studies, several estimators of the effects on road safety of various safety measures have been developed. The effects of daytime running lights have been assessed in terms of the accident rate ratio, the odds ratio and the ratio of odds ratios. Technical descriptions of these estimators of effect are given in the report presenting the meta-analysis of studies that have evaluated the effects on road safety of daytime running lights (Elvik, Christensen and Fjeld Olsen 2003). For the purpose of this paper, the question to be asked in a sensitivity analysis is: Does the choice of estimator of effect influence the results of the analysis?

The results of a meta-analysis ideally represent the combined contributions to knowledge of all the studies that are included in the analysis. It is therefore generally regarded as a problem if a single study exerts a decisive influence on the summary estimates of effect in a meta-analysis. One may refer to such cases as outlier bias. Testing for outlier bias should be a part of every sensitivity analysis of a meta-analysis.

Depending on the presence of between-study variation in estimates of effect often referred to as heterogeneity, different models of analysis may be adopted. Does the choice of a fixed-effects or random-effects model of analysis influence results? Would the results of a meta-analysis be different if all studies were weighted equally? These are questions a sensitivity analysis should answer.

The treatment of varying study quality in meta-analysis is a controversial topic (Christensen 2003). Different approaches have been proposed, ranging from not attempting to assess study quality at all to developing elaborate quality scores. In this paper, the position is taken that any meta-analysis should include an assessment of study quality and that one should test how sensitive the results of the meta-analysis are to study quality assessment.

To summarise, the items that should be subject to sensitivity analysis include:

1. The possible presence of publication bias
2. Choice of estimator of effect (if there is a choice)
3. The possible presence of outlier bias
4. Statistical weighting of studies included in a meta-analysis
5. Assessment of study quality

3. How can the results of sensitivity analysis be summarised?

What do we mean when saying that an analytic choice "makes a difference" for the results of a meta-analysis? In general, the results of a meta-analysis are presented as summary estimates of effect, associated with a confidence interval. As an example, one of the results of the meta-analysis of studies that have evaluated the effects of daytime running lights was this:

11% reduction of accidents (95% confidence interval: -14%, -8%)

Another result of the meta-analysis was this:

5% reduction of accidents (95% confidence interval: -7%, -3%)

These two estimates are based on different estimators of effect. The first estimate is based on the accident rate ratio; the second is based on the ratio of odds ratios. The two estimates are consistent, in the sense that both of them show a reduction of the number of accidents, and both of them are statistically significant at the 5% level. The two estimates do, however, differ with respect to the magnitude of the effect, and the confidence intervals do not overlap. One of the estimates is more precise than the other, having a 95% confidence interval spanning four percentage points, compared to six percentage points for the less precise estimate.

Sensitivity analysis proceeds by making a set of pair-wise comparisons of summary estimates of effect. In each such comparison, the outcome can be assessed in terms of three criteria:

1. The consistency of the estimates with respect to direction
2. The consistency of the estimates with respect to magnitude
3. The consistency of the estimates with respect to direction, magnitude and precision

In the above example, the two estimates are consistent with respect to the direction of impact. They are inconsistent with respect to magnitude of impact (the two confidence intervals do not overlap). The consistency of different summary estimates with respect to direction, magnitude and precision is assessed according to the degree of overlap of their confidence intervals. In the above example, there is no overlap.

To illustrate the use of these criteria, consider the following three 95% confidence intervals for summary estimates of effect:

Interval 1: (-19, +7); spanning 26 percentage points

Interval 2: (-11, -3); spanning 8 percentage points

Interval 3: (-22, -7); spanning 15 percentage points

Three pair-wise comparisons (1 vs 2, 1 vs 3, 2 vs 3) can be made to evaluate sensitivity on the basis of these confidence intervals. Comparing confidence intervals 1 and 2, it can be seen that they are consistent with respect to direction and magnitude of impact. The smaller confidence interval (2) is entirely contained within the larger (1). Confidence intervals 1 and 3 are also consistent with respect to direction and magnitude of impact, but do not fully overlap. A robustness score can be assigned as a function of the degree of overlap between two confidence intervals. Interval 1 is the larger of the two confidence intervals. If interval 3 had been contained entirely within interval 1, there would have been a full overlap, yielding a robustness score of 1. Interval 3 overlaps the range from –19 to –

7 of interval 1. This is an overlap of 12 percentage points out of 26 (the length of confidence interval 1), yielding a robustness score of 12/26 = 0.46. The robustness score is computed using the length of the larger of two confidence intervals as the denominator, and the share of this confidence interval covered by an overlapping, but smaller, confidence interval as the numerator. If two confidence intervals do not overlap at all, the robustness score is 0. If they are identical, the robustness score is 1. If one confidence interval is entirely contained within another, the robustness score is also 1, since such an overlap does not indicate any difference between two estimates in terms of direction or magnitude, merely in terms of precision. Since precise summary estimates are, ceteris paribus, preferred to less precise summary estimates, it seems reasonable to assign a robustness score of 1 to the case where a smaller confidence interval is entirely contained within a larger.

The results of a sensitivity analysis can thus be summarised in terms of three descriptors:

1. Consistency of summary estimates with respect to direction, described in terms of the number of consistent and number of inconsistent estimates for a set of N pair-wise comparisons.
2. Consistency of summary estimates with respect to magnitude, described in terms of the number of consistent and number of inconsistent estimates for a set of N pair-wise comparisons.
3. Robustness score, given as a number between 0 and 1, computed as the degree of overlap between two confidence intervals for a summary estimate.

It is proposed that the robustness scores can be classified as follows:

0.80-1.00 = Very robust
0.60-0.79 = Robust
0.40-0.59 = Neutral
0.20-0.39 = Sensitive
0.00-0.19 = Very sensitive

One would ideally want the results of a meta-analysis to be robust or very robust.

4. A case illustration of a sensitivity analysis: daytime running lights for cars

A meta-analysis has been reported for studies that have evaluated the effects on road safety of daytime running lights for cars (Elvik, Christensen and Fjeld Olsen 2003). The analysis included 25 studies. 13 of these have evaluated the intrinsic effects of daytime running lights, 12 have evaluated the aggregate effects of daytime running lights. By intrinsic effects are meant the effects on the accident rate of each car using daytime running lights, ideally speaking compared to an identical car not using daytime running lights. Aggregate effects refer to the effects of laws or campaigns that result in an increased rate of use of daytime running lights in a country or region, from, say, 40% use to 80% use. Estimates of intrinsic effects and estimates of aggregate effects are not directly comparable. They are therefore treated separately in the following analysis.

4.1. Sensitivity analysis of estimates of intrinsic effects of daytime running lights

As the first step in the sensitivity analysis, a funnel plot of estimates of the intrinsic effects of daytime running lights was prepared. This plot is shown in Figure 1. It shows 69 estimates of the intrinsic effects of daytime running lights, based on the estimator of effect (either the accident rate ratio, the odds ratio, or the ratio of odds ratios) preferred by the authors of each study. Each estimate of effect is plotted on the abscissa (using a log scale), a measure of its precision is plotted on the ordinate. In Figure 1, the fixed-effects statistical weight assigned to each estimate of effect is plotted on the ordinate. The reason why the

number of estimates of effect (69) is much greater than the number of studies (13) is that many studies have evaluated the effects of daytime running lights for several categories of accidents, providing one estimate for each category.

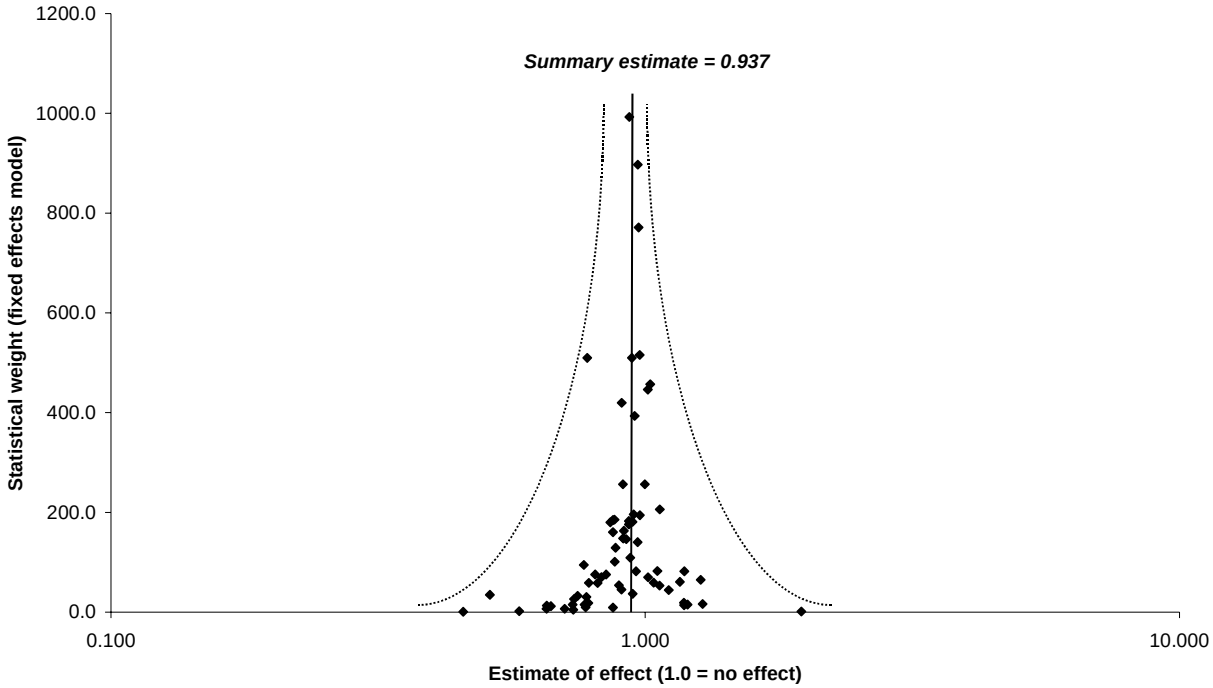

Figure 1: Funnel plot of estimates of the intrinsic effects of daytime running lights for cars

The idea underlying a funnel plot is that if all studies provide independent estimates of the same underlying mean effect, the scatter plot of estimates should resemble a funnel turned upside down. Contours have been added to Figure 1, clearly showing a funnel-like shape. Various techniques have been developed for analysing funnel plots in order to determine the possible presence of publication bias. It is beyond the scope of this paper to describe these techniques. For a concise description and discussion, see Christensen (2003).

The trim-and-fill technique developed by Duval and Tweedie (2000A, 2000B) was applied to Figure 1. This resulted in the addition of 9 data points according to the L-estimator of Duval and Tweedie. The addition of these data points had almost no impact on the summary estimates of effect. The robustness score for the fixed-effects summary estimate of effect (reported in Figure 1) was 0.945. The robustness score for the random-effects summary estimate of effect was 0.925. The overall robustness score with respect to publication bias was 0.935 (the average of the fixed-effects and random-effects scores).

Sensitivity was next tested with respect to the three estimators of effect applied in evaluation studies: the accident rate ratio, the odds ratio and the ratio of odds ratios. Summary estimates based on each of these estimators were available for five different categories of accidents, making it possible to conduct 15 (5 x 3) pair-wise comparisons. The results of these comparisons indicated that summary estimates of effects are indeed somewhat sensitive to the choice of estimator. Table 1 reports the results of the analysis. The overall robustness score was 0.540, which is clearly lower than desirable. This finding indicates that, rather than remaining agnostic about which estimator of effect to use, a choice should be made with respect to which of the three estimators of effect summary estimates should be based on. This issue will be further discussed in a subsequent section.

Table 1: Summary of results of sensitivity analysis of summary estimates of intrinsic effects of daytime running lights for cars

Level of analysis	Factor tested	Model of analysis or estimator of effect	Number of comparisons	Consistent direction	Inconsistent direction	Consistent magnitude	Inconsistent magnitude	Component robustness score	Overall robustness score	Overall rating
Intrinsic	Publication bias	Fixed-effects	2	2	0	2	0	0.945	0.935	Very robust
		Random-effects	2	2	0	2	0	0.925		
	Estimator	Fixed-effects	15	11	4	10	5	0.537	0.540	Neutral
		Random-effects	15	12	3	12	3	0.542		
	Outliers	Accident rate ratio	13	13	0	13	0	0.936	0.892	Very robust
		Odds ratio	13	12	1	12	1	0.846		
		Ratio of odds ratios	13	13	0	13	0	0.895		
	Weighting	Accident rate ratio	3	3	0	3	0	1.000	0.700	Robust
		Odds ratio	3	1	2	1	2	0.200		
		Ratio of odds ratios	3	3	0	3	0	0.895		
	Quality	Fixed-effects	3	3	0	3	0	0.573	0.702	Robust
		Random-effects	3	3	0	3	0	0.830		

The next item tested in the sensitivity analysis was the possible presence of outlier bias, or more precisely, a single study exerting a decisive influence on the summary estimate of effect. This testing was done by omitting one study from the data set and estimating a summary estimate of effect based on the remaining N-1 studies. Since there were 13 studies of the intrinsic effects of daytime running lights, this procedure was repeated 13 times for each of the three estimators of effect. The summary estimates based on N-1 studies were compared to the one based on all N studies to determine any differences. The procedure was performed for random-effects estimates of effect only, as there was significant between-study variation in estimates of effect.

Despite the use of a random-effects model, an outlying study was found for the odds ratio estimator of effect (see Table 1). This was a fairly large study whose findings were highly inconsistent with those of the remaining 12 studies. For the other two estimators of effect, the accident rate ratio and the ratio of odds ratios, no outlying studies were found.

As far as statistical weighting is concerned, three weighting schemes were compared: (1) Fixed-effects weights, (2) Random-effects weights, and (3) No weights, i.e. giving all studies the same weight. Choice of statistical weights was found to influence summary

estimates of effect for the odds ratio estimator, but not for the other two estimators of effect (see Table 1). Overall robustness score was 0.700.

Finally, the assessment of study quality was considered. Studies were formally scored for quality, and a quality score ranging from 0 to 1 was assigned to each study (Elvik, Christensen and Fjeld Olsen 2003). Figure 2 shows the relationship between quality score and the overall estimate of effect for each of the 13 studies that have evaluated the intrinsic effects of daytime running lights. The overall estimate of effect for each study was simply a weighted sum of the estimates for each category of accident, based on the preferred estimator of effect in each study.

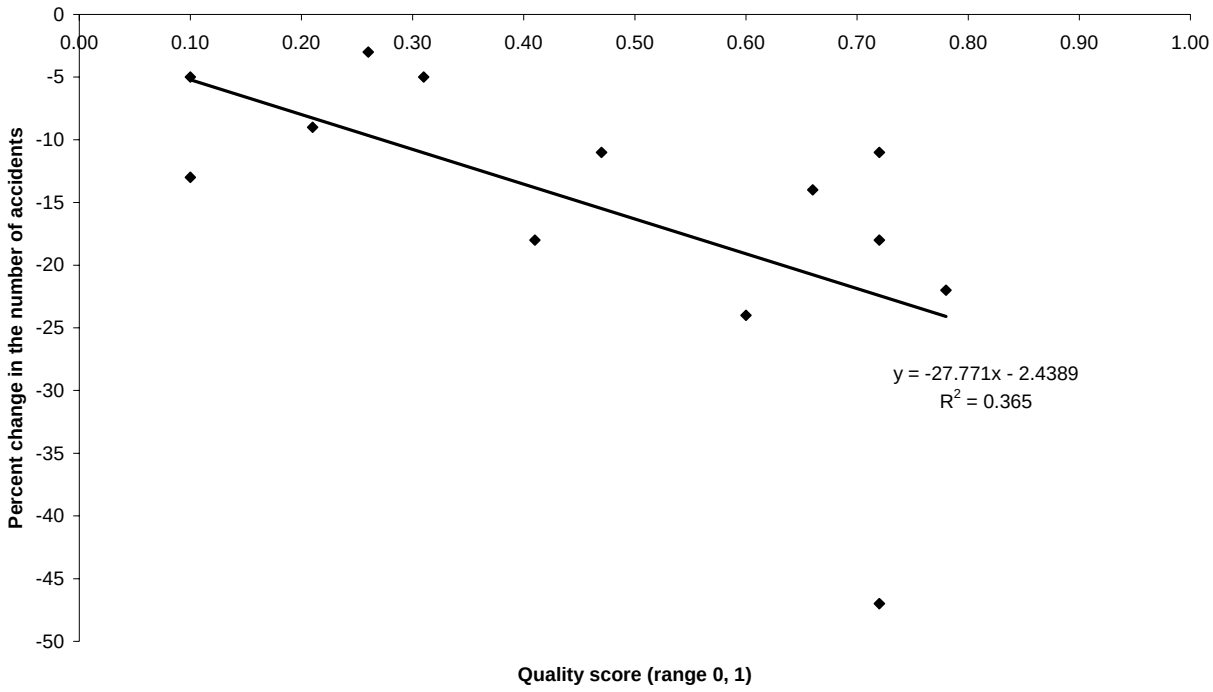

Figure 2: Scatter plot of study quality score versus estimate of the intrinsic effects of daytime running lights for cars

A straight line has been fitted to the data points in figure 2, showing that the higher the quality score for a study, the larger is the accident reduction attributed to daytime running lights by that study. This tendency runs counter to the "Stainless Steel Law of Evaluation", one of the metallic laws for evaluation studies proposed by Rossi (1987). Besides, the dispersion of estimates of effect appears to increase as study quality improves, which is also somewhat contrary to what one would expect. At any rate, there is a statistically significant relationship between study quality and estimate of effect (Pearson's r = .604, $F_{1,11} = 6.32$, p = 0.029), which shows that the assessment of study quality could influence summary estimates of effect in the meta-analysis.

In the meta-analysis, adjusted estimates of effect were estimated by multiplying statistical weight by quality score, thus effectively reducing the weight given to poor studies more than the weight given to good studies. Bérard and Bravo (1998) applied a similar technique in a meta-analysis of studies evaluating bone loss in postmenopausal women. However, as pointed out by Christensen (2003), their application of the technique was erroneous for the random-effects model. Christensen provides a correct formula for the random-effects model, which was applied here.

As shown in Table 1, adjusting for study quality hardly influenced the results of the meta-analysis, in particular not for the random-effects analysis. Overall robustness score with respect to adjustment for study quality was 0.702.

4.2. Sensitivity analysis of estimates of aggregate effects of daytime running lights

The sensitivity analysis of estimates of the aggregate effects of daytime running lights for cars proceeded the same way as the sensitivity analysis of estimates of the intrinsic effects. Each step of the analysis will therefore not be described, but the results will be presented and commented.

Figure 3 shows a funnel plot of estimates of the aggregate effects of daytime running lights. There are 42 data points based on 12 studies. Several studies estimated effects for more than one category of accidents; hence the number of estimates of effect exceeds the number of studies.

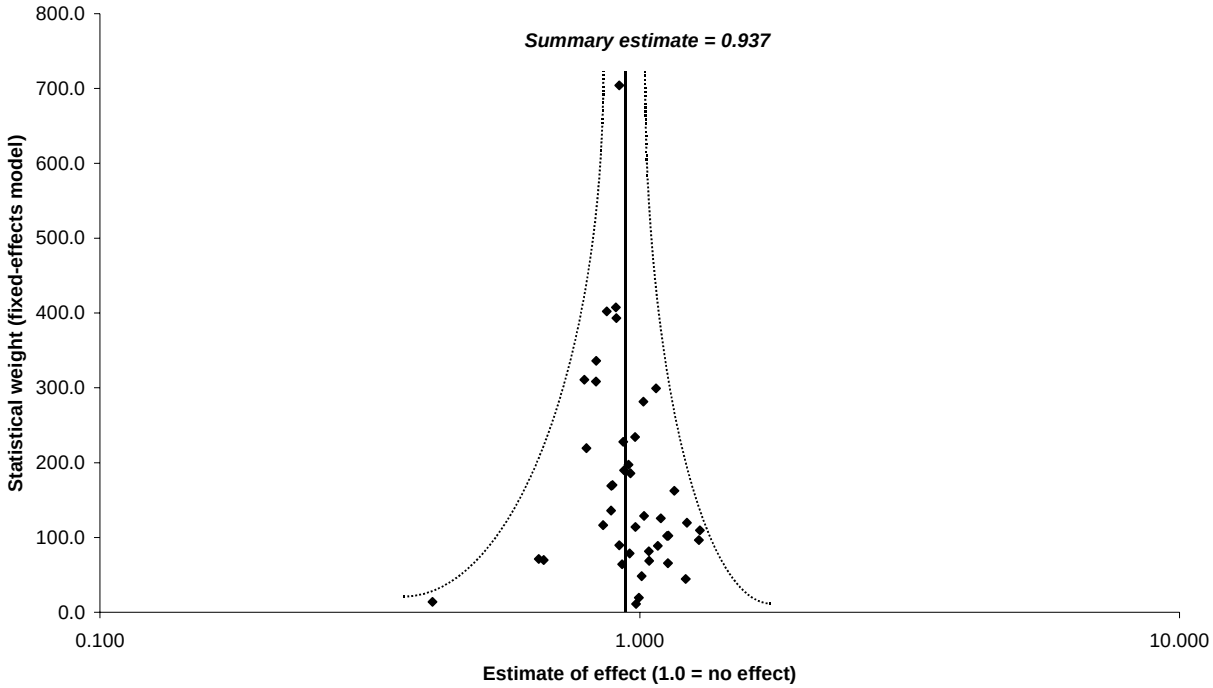

Figure 3: Funnel plot of estimates of the aggregate effects of daytime running lights for cars

Figure 3 looks rather similar to Figure 1, which showed estimates of the intrinsic effects of daytime running lights. Application of the trim-and-fill technique resulted in the addition of 2 data points based on the R-estimator of Duval and Tweedie, but the addition of these new data points had almost no effect on summary estimates of effect. As shown in Table 2, the overall robustness score with respect to publication bias was 0.920.

A total of 21 comparisons were made to evaluate sensitivity to the choice of estimator of effect. The results were not very encouraging as far as the fixed-effects model is concerned. Several results were inconsistent with respect to magnitude, and the robustness score was only 0.242 (see Table 2). The random-effects model fared a lot better, getting a robustness score of 0.679.

No outliers were found (see Table 2). The overall robustness score with respect to the presence of outliers was 0.909. Choice of statistical weighting did not affect summary

estimates of effect. Overall robustness score was 0.880. Finally, adjusting for study quality did not influence summary estimates of effect. The robustness score with respect to adjustment for study quality was 0.963 (see Table 2).

Table 2: Summary of results of sensitivity analysis of summary estimates of aggregate effects of daytime running lights for cars

Level of analysis	Factor tested	Model of analysis or estimator of effect	Number of comparisons	Consistent direction	Inconsistent direction	Consistent magnitude	Inconsistent magnitude	Component robustness score	Overall robustness score	Overall rating
Intrinsic	Publication bias	Fixed-effects	2	2	0	2	0	0.930	0.920	Very robust
		Random-effects	2	2	0	2	0	0.910		
	Estimator	Fixed-effects	21	21	0	10	11	0.242	0.460	Neutral
		Random-effects	21	21	0	21	0	0.679		
	Outliers	Accident rate ratio	12	12	0	12	0	0.898	0.909	Very robust
		Odds ratio	12	12	0	12	0	0.898		
		Ratio of odds ratios	12	12	0	12	0	0.908		
	Weighting	Accident rate ratio	3	3	0	3	0	0.857	0.880	Very robust
		Odds ratio	3	3	0	3	0	1.000		
		Ratio of odds ratios	3	3	0	3	0	0.780		
	Quality	Fixed-effects	3	3	0	3	0	0.927	0.963	Very robust
		Random-effects	3	3	0	3	0	1.000		

Figure 4 shows the relationship between study quality and estimate of effect for the twelve studies that have evaluated the aggregate effects of daytime running lights for cars. In this case, the relationship between study quality and estimate of effect is somewhat weaker than for studies that have evaluated the intrinsic effects of daytime running lights. A tendency is also seen for estimates of effect based on poor studies to be more widely dispersed than estimates of effect based on good studies.

The main findings of the sensitivity analysis can be summarised as follows:

1. Publication bias: There was no evidence of any great publication bias. Estimates of effect were not sensitive to adjusting for publication bias. The robustness score exceeded the value of 0.9 both for estimates of intrinsic effects and estimates of aggregate effects.

2. Choice of estimator of effect: Estimates were found to be sensitive to the choice of estimator of effect, in particular for the fixed-effects model of analysis applied to estimates of aggregate effects.

3. Outlier bias: One outlying study was identified for the odds ratio estimator of intrinsic effects. Otherwise, no outlying studies were identified.

4. Statistical weighting: Summary estimates of effect were, in general, not found to be sensitive to how individual estimates were weighted in the meta-analysis.

5. Adjusting for study quality: Summary estimates of effect were not found to be sensitive to adjustment for study quality.

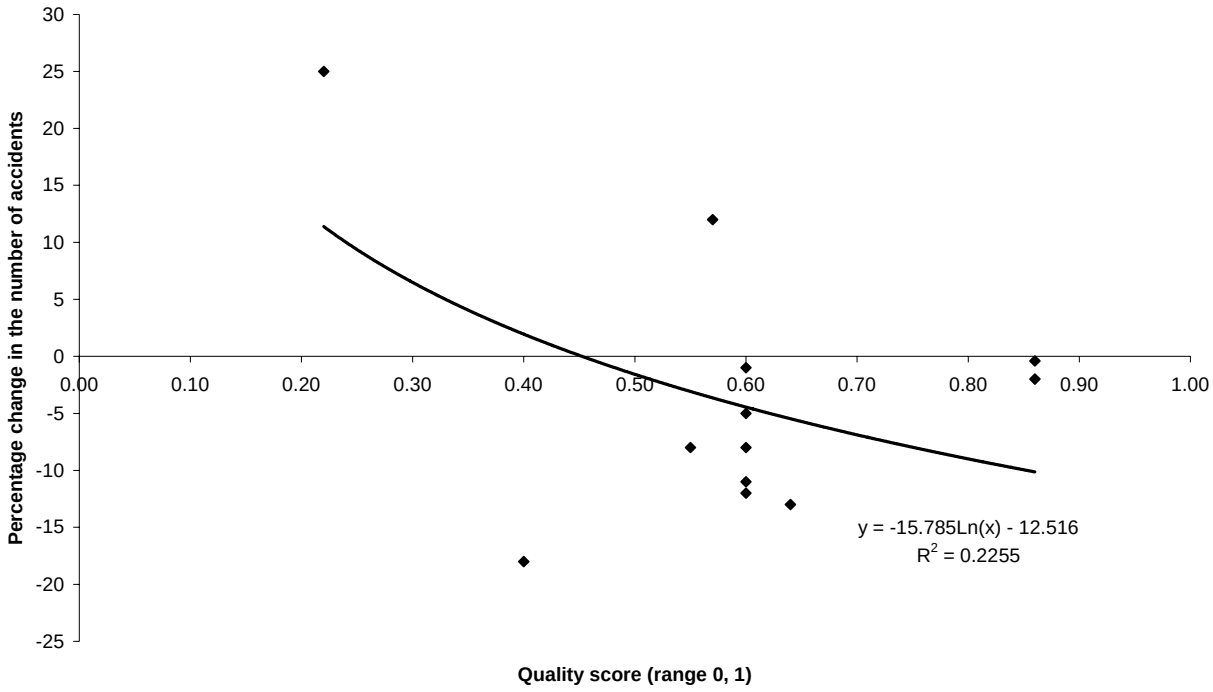

Figure 4: Scatter plot of study quality score versus estimate of the aggregate effects of daytime running lights for cars

5. How can the results of a sensitivity analysis guide analytic choices?

Can a sensitivity analysis in any way guide the analytic choices made in a meta-analysis? Yes and no. Yes, in the sense that such an analysis sheds lights on the effects of these choices with respect to the results of the analysis. No, in the sense that the meta-analysis has to be performed before the consequences of the choices made as part the analysis are seen. Hence, a sensitivity analysis cannot guide analytic choices before they are made, but it can, by shedding light on the consequences of those choices, show their implications for the results of the meta-analysis and for the interpretation of those results.

Returning to the sensitivity analysis of the meta-analysis of studies that have evaluated the effects on road safety of daytime running lights for cars, it is, in the first place, obvious that in presenting these results, it is not necessary to adjust them for publication bias, or interpret them conservatively out of fear that unpublished studies are likely to exert a large influence on summary estimates of effect. As far as the statistical tests for publication bias go – and all these tests are admittedly based on various assumptions that are by themselves not testable – there are no indications that any important publication bias is present. The

results are robust with respect to this potential source of error. This should increase confidence in the findings of the meta-analysis.

In the second place, the sensitivity analysis found that the choice of estimator of effect does indeed matter in this case. The question has to be asked: Which of the three estimators should be preferred? The odds ratio estimator of effect should be rejected, because it was found to contain an outlying estimate of effect that strongly influenced the summary estimate. As far as the remaining two estimators of effect are concerned, the one that gives the most precise summary estimates of effect should be preferred (assuming, of course, that this is not "spurious precision" generated by large studies of poor quality). If a random-effects model is assumed, the ratio of odds ratios estimator of effect should be preferred.

In the third place, a random-effects model is more robust than a fixed-effects model. This is hardly a surprising finding, given the fact that the data set contained significant between-study variation in estimates of effect. Giving equal weight to all estimates of effect does not minimise the variance of the summary estimate of effect, and is therefore a weighting scheme that should be rejected. Despite the widespread agreement on this among statisticians working on techniques for meta-analysis, studies that give equal weight to all individual estimates of effect have been reported (Wagenaar et al 1995).

In the fourth place, adjusting studies for quality does not influence summary estimates of effect. This is a reassuring finding, for at least two reasons. The first reason is that there is no consensus on how best to treat study quality in meta-analysis. Opinions range from ignoring the issue, arguing that any assessment of study quality is subjective, to including an adjustment for study quality based on a quality score, as was done in the sensitivity analysis reported here. The second reason why robustness with respect to study quality is reassuring, is that if this had not been the case – if results had been highly sensitive to study quality – a decision would have had to be made about throwing out bad studies. This would have introduced an arbitrary threshold for study quality, and might have involved throwing away a sizable proportion of studies.

In summary, a sensitivity analysis can guide the analyst in presenting the results of a meta-analysis. In the present case, the sensitivity analysis clearly indicates that the most credible summary estimates are those that are based on the ratio of odds ratios estimator of effect and a random-effects model of analysis. No adjustment for publication bias or study quality is needed.

6. Discussion

This paper has argued that sensitivity analyses of meta-analyses should be based on a systematic approach, just as any other element of meta-analysis. There are three key elements of the systematic approach proposed in this paper:

1. A specification of what a sensitivity analysis should include. It should include: publication bias, choice of estimator of effect, outlier bias, choice of statistical weights, and assessment of study quality.
2. A specification of the order in which the items considered should be analysed: start by testing for publication bias; continue by testing for choice of estimator of effect, outlier bias and statistical weighting, and end by testing for study quality assessment.
3. A proposal for how to describe the results of a sensitivity analysis, by stating how consistent summary estimates are with respect to direction and magnitude of effect, and by estimating a robustness score, ranging from 0 to 1.

There are no doubt other sources of error in meta-analysis than those considered here. The items listed above should be seen as mandatory: Any sensitivity analysis should consider these items, but may consider other items as well.

The reason it is proposed to start the analysis by testing for publication bias is that this is a potentially very serious source of error in meta-analysis. If there is evidence of severe publication bias, the results of the meta-analysis are simply misleading. A meta-analysis affected by severe publication bias should either not be presented at all, or some attempt should be made to adjust for the bias. In principle, the trim-and-fill technique adjusts for publication bias. It is, however, based on assumptions regarding the distribution of study findings (an assumption of symmetry around the summary mean) that may not always be correct. It is therefore best applied as a method for sensitivity analysis only, to indicate the magnitude of the effect of publication bias, and not to actually adjust summary estimates for the purpose of removing the bias.

A similar point of view applies to the other sources of error in meta-analysis. Suppose, as an example, that there is outlier bias. Should the outlying study be omitted? Not necessarily. What looks like bias, could have a perfectly legitimate explanation. It could be the case, for example, that a certain treatment has a favourable effect in some groups of patients, but an adverse impact in other groups of patients. The deviant study may have used an entirely different sample of patients than did the other studies. The issue then rather becomes that of mixing apples and oranges, not of throwing out studies that somehow went awry. One would, however, have to go fairly deeply into the details of the studies in order to know whether a deviant finding is real or an artefact of study method.

This paper has argued that a sensitivity analysis may guide – albeit after the fact – some of the analytic choices made as part of a meta-analysis. As the above example suggests, the guidance that can be extracted from a sensitivity analysis is sometimes limited. A sensitivity analysis may alert the analyst to choices that must be made, but does tell him what the right choice is. This difficulty is compounded if it turns out that a meta-analysis is quite sensitive to several of the tested items. Should the analyst then both try to adjust for publication bias, throw out outlying studies, and adjust for study quality? Most researchers would probably agree that he should not try to manipulate the meta-analysis to set right all these sources of error. The most sensible thing to do in case a meta-analysis is found to be sensitive to multiple choices made by the analyst is not to adjust the results of the analysis, since that would inevitably involve a large element of subjectivity. If a meta-analysis is found to be sensitive to the items tested, it is rather more informative to tell readers, as precisely as possible, what this implies for the results of the analysis and their interpretation.

7. Conclusions

The main conclusions of the research reported in this paper can be summarised as follows:

1. A systematic approach to sensitivity analysis in meta-analysis has been proposed. Key elements of the approach include a specification of the items to be included in a sensitivity analysis, a specification of the order in which sensitivity to these items should be tested, and a proposal for how to describe the results of a sensitivity analysis.
2. A new index, the robustness score, is proposed. It is derived from the degree of overlap between two confidence intervals for summary estimates, which are compared as part of a sensitivity analysis. The robustness score ranges from 0 to 1.
3. An example of a sensitivity analysis is given, based on a recently reported meta-analysis of studies that have evaluated the road safety effects of daytime running

lights for cars. This analysis found that summary estimates were sensitive to the choice of estimator of effect, but otherwise robust.

4. In principle, a sensitivity analysis can guide some the analytic choices that are subject to a sensitivity analysis, albeit after the fact. The main use of a sensitivity analysis is, however, to help users of meta-analyses assess how much confidence they can place in the results of such analyses, in view of the fact that there are many ways of performing meta-analyses.

References

Bérard, A., Bravo, G. 1998. Combining studies using effect sizes and quality scores: Application to bone loss in postmenopausal women. Journal of Clinical Epidemiology, 51, 801-807, .

Christensen, P., 2003. Topics in meta-analysis. A literature survey. TOI report 692. Oslo, Institute of Transport Economics.

Deeks, J. J., 2002. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. Statistics in Medicine, 21, 1575-1600.

Duval, S., Tweedie, R. 2000A. A non-parametric "Trim-and-fill" method of accounting for publication bias in meta-analysis. Journal of the American Statistical Association, 95, 89-98.

Duval, S., Tweedie, R., 2000B. Trim-and-fill: A simple funnel plot based method of testing and adjusting for publication bias in meta-analysis. Biometrics, 56, 276-284.

Elvik, R., Christensen, P., Fjeld Olsen, S., 2003. Daytime running lights. A systematic review of effects on road safety. TOI report 688. Oslo, Institute of Transport Economics.

Rossi, P. H., 1987. The Iron Law of evaluation and other metallic rules. Research in Social Problems and Public Policy, 4, 3-20.

Wagenaar, A. C., Zobeck, T. S., Williams, G. D., Hingson, R. D., 1995. Methods used in studies of drink-drive control efforts: a meta-analysis of the literature from 1960 to 1991. Accident Analysis and Prevention, 27, 307-316.