Scand J Work Environ Health 2012;38(3):282-290
https://doi.org/10.5271/sjweh.3201
Published online: 21 Oct 2011, Issue date: May 2012

Synthesizing study results in a systematic review

by Verbeek J, Ruotsalainen J, Hoving JL

Affiliation: Finnish Institute of Occupational Health, Cochrane Occupational Safety and Health Review Group, PO Box 310, 70101 Kuopio, Finland. jos.verbeek@ttl.fi

Refers to the following texts of the Journal: 1995;21(2):134-142 2001;27(6):388-394 2001;27(4):258-267 2002;28(6):386-393 2002;28(5):314-323 2011;37(1):1-5 2007;33(2):81-83

The following articles refer to this text: 2013;39(6):633-634; 2014;40(2):133-145; 2014;40(3):215-229; 2018;44(2):134-146

Key terms: clinical heterogeneity; intervention; knowledge translation; levels of evidence; meta-analysis; occupational health; review; systematic review

This article in PubMed: www.ncbi.nlm.nih.gov/pubmed/22015561

Print ISSN: 0355-3140 Electronic ISSN: 1795-990X

Discussion paper

Scand J Work Environ Health 2012;38(3):282–290. doi:10.5271/sjweh.3201

Synthesizing study results in a systematic review

by Jos Verbeek, MD, PhD,[1, 2] Jani Ruotsalainen, MSc,[1] Jan L Hoving, PhD [2]

Verbeek J, Ruotsalainen J, Hoving JL. Synthesizing study results in a systematic review. *Scand J Work Environ Health*. 2012;38(3):282–290 doi:10.5271/sjweh.3201

A single study rarely suffices to underpin treatment or policy decisions. This creates a strong imperative for systematic reviews. Authors of reviews need a method to synthesize the results of several studies, regardless of whether or which statistical method is used. In this article, we provide arguments for combining studies in a review. To combine studies, authors should judge the similarity of studies. This judgement should be based on the working mechanism of the intervention or exposure. It should also be assessed if this mechanism is similar for various populations and follow-up times. The same judgement applies to the control interventions. Similar studies can be combined in either a meta-analysis or narrative synthesis. Other methods such as vote counting, levels of evidence synthesis, or best evidence synthesis are better avoided because they may produce biased results. We support our arguments by re-analysing a systematic review. In its original form, the review showed strong evidence of no effect, but our re-analysis concluded there was evidence of an effect. We provide a flow-chart to guide authors through the synthesis and assessment process.

Key terms clinical heterogeneity; intervention; knowledge translation; levels of evidence; meta-analysis; occupational health.

The basic idea underlying evidence-based medicine is that better use of evidence from scientific research will increase the quality of healthcare including prevention (1). Evidence is, however, seldom unequivocal and many topics of interest to practitioners have been evaluated in more than one study with varying results. This creates a clear need for synthesizing the results of multiple studies such as in systematic reviews. The systematic review has been defined as a review in which bias has been reduced by the systematic identification, appraisal, synthesis, and, if relevant statistical aggregation of all relevant studies on a specific topic according to a predetermined and explicit method (2). The value of systematic reviews in providing answers to questions relevant to practice is increasingly recognized also for occupational health (3, 4).

In the past, rather than providing an answer to a specific question, the purpose of a review was to give an overview of what had been written about a certain topic in the scientific literature. For this traditional "overview type" of review, synthesis of the results in one summary outcome is less necessary. This difference in objectives has created confusion about if, when, and how results of studies in reviews should be synthesized.

Not all types of questions can be answered with systematic reviews. The traditional idea of giving an overview of "the state of the art" can still be useful. It is, however, increasingly recognized that also in this respect it would be good to be more systematic. This has led to a new nomenclature for reviews such as "scoping" reviews (5). The objective of a scoping review is to summarize a range of evidence in order to convey the breadth and depth of a field. Such reviews have requirements different than systematic reviews as defined above. Results of qualitative studies can also be combined in a synthesis of studies, but the problems here are different from those in quantitative studies (6). Therefore, in this article, we restrict ourselves to systematic reviews of quantitative studies only.

There is sometimes confusion about the difference between a systematic review and a meta-analysis. A systematic review is a review of the literature, but it

1 Finnish Institute of Occupational Health, Cochrane Occupational Safety and Health Review Group, Kuopio, Finland.
2 Coronel Institute of Occupational Health and Research Center for Insurance Medicine, AMC-UMCG-UWV-VUMC, Academic Medical Center, Amsterdam, the Netherlands.

Correspondence to: Jos Verbeek, MD, PhD, Finnish Institute of Occupational Health, Cochrane Occupational Safety and Health Review Group, PO Box 310, 70101 Kuopio, Finland. [E-mail: jos.verbeek@ttl.fi]

does not necessarily include a meta-analysis. A meta-analysis is a statistical synthesis of the results of several individual studies in one pooled summary estimate. As such, it is easy to see that a meta-analysis requires a systematic review of the literature. Since a meta-analysis is often included in a systematic review, many use the term meta-analysis as a synonym for systematic review (7). Meta-analysis has a long history in educational and psychological research (8). The statistical technique of combining study results is not difficult and the pooled effect estimate has the charm of simplicity. However, this pooled effect estimate does not have much meaning if this comes from primary studies that widely vary in types of exposures, interventions, or participants. Meta-analysis has therefore been criticized for comparing apples to pears, and authors have been cautioned against combining study results too easily (9).

Regardless of whether a statistical method is used or not, authors will always need a method to synthesize the results of several studies to be able to provide answers to practical questions. The challenge will be to strive for a valid answer that is as concise and succinct as possible. For interventions, we would ultimately like to know how well the intervention works, and for exposures we would like to know to what degree they cause ill-health. The method of combining study results is not a trivial problem as the results of reviews can widely vary depending on the type of study synthesis used. Moreover, the validity of a systematic review has more direct practical implications than a primary study as its results are more likely to be used for policy making or to underpin clinical practice guidelines than the results of a single study.

Therefore, we would like to provide an overview of methods for synthesizing study results in a systematic review and assess their pros and cons.

The review process and decisions on study synthesis

The question how to synthesize study results is important from the very inception of a systematic review. During the process of performing a systematic review, several steps are taken that influence the synthesis of individual studies. In the first phase of the review, during the operationalization of the inclusion criteria, the author must determine whether similar or dissimilar studies are included and thus whether studies can be combined or not (2). How studies can be combined is a problem we will address later. The more important question is which studies have sufficient similarity to give an interpretable pooled estimate of the effect of the intervention or exposure. This will always be a subjective assessment because, in the end, no two studies will be identical. It is self-evident that the similarity of the studies depends on the inclusion criteria of the review. Sometimes, authors have formulated these criteria so broadly that studies can never be sensibly combined unless they are divided into different categories. They state, for example, that they want to study the effect of "interventions" on a certain health problem, meaning that they want to include all possible interventions. Or they state that they want to study the effect of a broad type of intervention such as “behavioral interventions” or "exercise". This is of course not impossible but it actually means that one performs several reviews under the umbrella of one review. This is not always recognized. Authors often state that the included studies were so heterogeneous that they could not be combined into a meta-analysis without noting that this was due to their own broad inclusion criteria (10–12). Instead of making subcategories and performing a meta-analysis, authors then still combine studies. For the study synthesis, they do not explain how they combined the results or they use a "self-invented" method for synthesis often leading to biased results of the systematic review. Ioannidis et al (13) has especially pointed this out, arguing that authors of reviews should better underpin their decisions about heterogeneity and more often make use of meta-analysis. This does not preclude the combination of broad clinical questions into a meta-analysis as shown in several systematic reviews, but there has to be proper argumentation to make the summary estimate credible (14, 15).

In the literature, the criteria for combining studies are often referred to as clinical and statistical heterogeneity. Clinical heterogeneity means that any feature of the included studies can be so divergent that it precludes synthesis. Clinical heterogeneity is not an intuitive concept because it is unclear what clinical means in this context. Therefore, we would prefer to use the word "similarity" of studies instead. Statistical heterogeneity is the variation in treatment effect that is due to differences between studies rather than by chance alone (16). Even though studies can be judged similar enough to be combined, the statistical heterogeneity can be so considerable that it does not make sense to combine the results. If, for example, some studies have a large beneficial effect and other studies have a harmful effect, then it does not make sense to combine the results and state that there is no effect. In that case, there are probably differences between the studies that we did not understand or cannot estimate with the data at hand (17).

How to judge if studies are similar?

In figure 1, we provide a flow-chart of the argumentation for combining or not combining studies. Our primary feature of interest in studies is usually the intervention or the exposure, and this should therefore be the point of departure. If the interventions are not similar, then the results of studies should be reported separately, either in separate systematic reviews or separate sections of one systematic review. We advocate judging the similarity of interven-

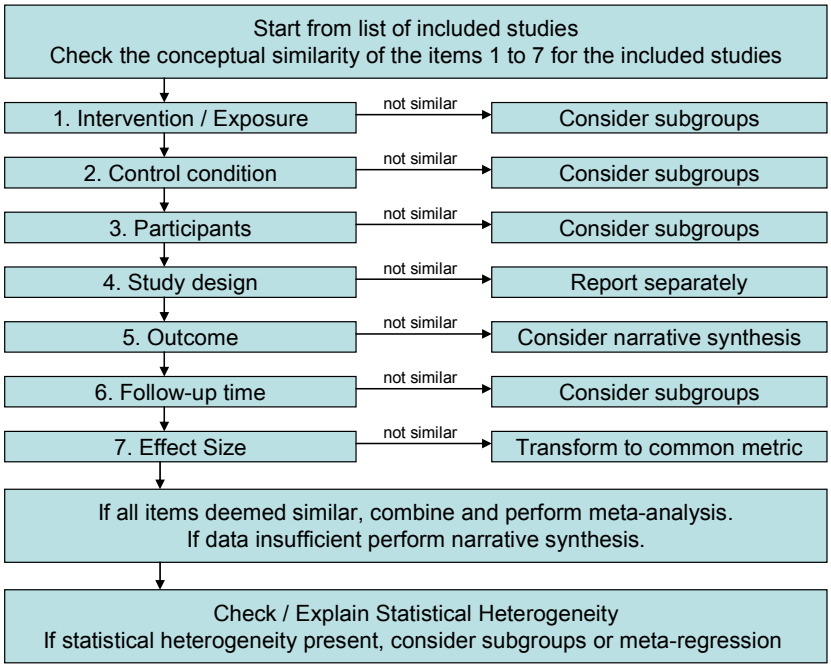

Figure 1. Flowchart for deciding about synthesis of study results in a systematic review.

tions by their mechanism or action based on which one would expect a similar effect of the intervention. This is a subjective judgment because a general intervention classification does not exist. It is not an easy judgment to make because the meaning of the intervention has to be interpreted by the authors of the systematic review based on the short description that is provided in the report of the primary study (18). With complex interventions such as behavioral or organizational changes, the judgment can especially be complicated. The intervention feature of interest can only be a small part of the whole intervention, and then it is unclear what is being combined. For example pedometers or step counters to increase physical activity are usually part of a broader package of measures to induce a less sedentary lifestyle. Other features that could be part of an intervention package, such as professional guidance in exercises or working time availability for exercise, can be more crucial and thus make interventions dissimilar (19). Recently, more emphasis has been put on the systematic development of interventions such as with intervention mapping or the use of logic models (20). In addition, articles that report protocols of randomized controlled trials allow more room for an extensive description of the intervention. These developments will enable better judgment of similar interventions.

As a second step, authors should assess the control condition because it is conceivable that a no-intervention control group will have a different effect than a less-intensive intervention control group. Here, at least the following types of control conditions can be discerned: no intervention, a waiting-list control condition that will get the intervention later, a true placebo or sham treatment, alternative interventions, and similar but less intensive interventions. Judgment of similarity depends also here on the mechanism by which the effect is brought about. If there would not be consensus about the working mechanism, then the effect of different working mechanisms on the conclusions should be examined in a sensitivity analysis.

Interventions can have a different effect on various participants, for example, children or adults (21). It can also be surmised that the intervention would work similarly in various occupations that are subject to the same exposure. For example, we expected the same effect on back pain of training in manual handling of patients and materials among nurses and baggage-handlers because the mechanism was deemed similar and judged to produce similar results (22).

In general, it is not recommendable to combine different study designs such as randomized and non-randomized studies (23). The idea is that different designs will lead to different types and degrees of bias and that therefore the summary estimate will be difficult to interpret.

Outcomes that are conceptually dissimilar should also not be combined even though it would be technically easy to do. For example the effect of reduction of exposure for treating occupational asthma can be measured on asthma symptoms and sick leave days due to asthma (24). It can be assumed that these effects would be different and cannot be combined. On the other hand, if the authors are interested in the effect of physical

conditioning on sick leave among back pain patients, it makes sense to combine time to return to work and the mean number of sick leave days as outcomes. Both types of outcomes measure the same concept, thus it can be assumed that the intervention has a similar effect on both types of outcomes (17).

For many interventions, such as educational interventions, it would be plausible that they have a differential effect over time. Sometimes there could be a learning period after which a full effect is expected or the effect could wear off over time. Depending on the mechanism that is anticipated, only outcomes at similar follow-up times should be combined. In our view, it does not make sense to split this into too small parts because then there will never be enough studies to combine. This is a specific problem for studies that use back pain as an outcome. Here, the experts expect a differential effect of intervention in the short term after three months follow-up, after a year follow-up, and after longer time periods. There is, however, no empirical evidence that this is a valid categorization (25).

Once the authors of the systematic review have decided whether the studies' elements are similar enough to be combined, they must then assess if the data in the original reporting is appropriate to enable a statistical meta-analysis. Statistically, it is only possible to combine study results that are measured in a similar way, such as dichotomous outcomes as odds ratios (OR) or rate ratios or continuous outcomes as mean differences. However, simple methods exist to transform effect-sizes for dichotomous outcomes into effect sizes for continuous outcomes and vice versa. This greatly facilitates the conduction of meta-analysis. We refer to Borenstein & Cooper for an extensive and didactic overview of these methods (7, 8).

Meta-analysis and statistical heterogeneity

After authors follow this procedure and have decided that studies may be combined and their data is appropriate, they can proceed with the meta-analysis. Software for meta-analysis is freely available from the Cochrane Collaboration if not used for commercial purposes (Review Manager 5.1, Cochrane Collaboration, Copenhagen, Denmark). Also other statistical programmes have sophisticated options for meta-analysis such as Stata version 9 (StataCorp, College Station, TX, USA). In a meta-analysis, the study results are weighted according to their precision or variance, where studies with greater precision get a higher weight. The pooled estimate is then calculated based on these weighted study effect sizes. The results can also be presented graphically in a forest plot which gives an immediate overview of the individual studies and their statistical heterogeneity (26). In figure 2, the weight of the studies is based on the standard error of the log OR, with more precise studies with smaller standard errors having more weight.

High statistical heterogeneity means that the between-study variance is higher than would be expected based on chance alone. When there is high statistical heterogeneity, this should be analyzed for example by dividing studies into different subgroups (27). Subgroups could show different pooled effect estimates and thus explain the heterogeneity in the whole sample of studies. Ultimately meta-regression can be used, where characteristics of studies are regressed on the effect sizes to find out if this explains effect-size variations. Since the sample sizes of studies included in a systematic review are usually small, the power of meta-regression is low. Therefore it is recommended to use this only as a hypothesis-generating technique (28).

Narrative synthesis

If a statistical combination of studies is not possible for example because the various study elements are not similar enough, the only alternative is a narrative synthesis. This is not essentially different from the procedure described above up to the point of statistically combining the results. Instead of combining them, the results are simply described as well as possible. This has been elaborated by Rodgers et al (30) for interventions to increase ownership of smoke alarms. The authors performed independently both a narrative synthesis and a meta-analysis for one particular systematic review. Their conclusion was that the final conclusions were similar but that a narrative synthesis provided more ideas for implications for future research and the meta-analysis more ideas for moderators of the effect of the intervention (29, 30).

Alternative synthesis methods

Vote counting. Vote counting is best described as summing up the numbers of studies with statistically significant outcomes and those without significant outcomes. If those with statistically significant outcomes prevail then it is concluded that there is evidence for the effectiveness of an intervention or exposure (31). The main argument against the vote counting method is that, for studies with low statistical power, the approach easily leads to the conclusion that there is no effect while in reality there is an effect.

Levels of evidence. Levels of evidence are best described by the Cochrane Back Review Group in their previous methods guidelines published in 2003 (32). The group followed the same approach as described above for

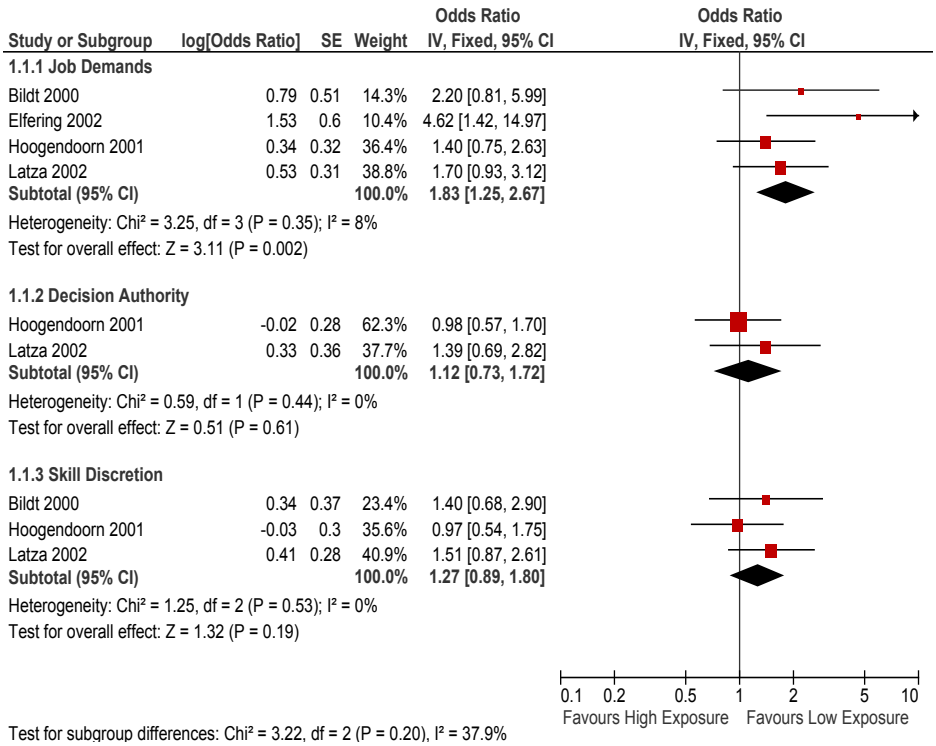

Figure 2. Results of re-analysis of studies in systematic review by Hartvigsen et al (40).

judging the clinical homogeneity. If studies are homogeneous, they are synthesized into a level of evidence for or against the effectiveness of an intervention. The summing depends on the quality of the studies and is summarized as strong, moderate, low, or conflicting evidence.[1] The levels of evidence method is sometimes also called "qualitative synthesis".

The levels of evidence synthesis should not be confused with the overall judgment of the quality of evidence in a systematic review as proposed by the Grading of Recommendations, Assessment, Development and Evaluations (GRADE) working group (23, 33). The GRADE method is an overall judgment of the quality of evidence and not a method to synthesize study results. In addition to a summary measure of effect, such as a pooled relative risk or OR, the quality of the evidence is rated as high, moderate, low, or very low. The working group advocates the use of the following five criteria to judge the quality of the evidence: (i) risk of bias in the included studies, (ii) indirectness of the evidence, (iii) unexplained heterogeneity of the results of the included studies, (iv) imprecision of the results, and (v) probability of publication bias. Thus the quality of the evidence reflects the confidence that the estimate of effect is correct.

One of the problems with the levels of evidence synthesis is the definition of a positive or negative outcome. A positive outcome is usually defined when there is a statistically significant positive outcome at a level of $P<0.05$ and a negative outcome if there is a non-significant outcome. A consistent finding would then be that four out of five trials had a significantly positive outcome.

The advantage of the levels of evidence synthesis is that it is saves a lot of work because no laborious data extraction is needed. One only has to know the P-value of the outcome to be able to combine study results. In addition, one can synthesize evidence for or against effectiveness. A serious drawback of the method is that its criteria are not well defined. Ferreira et al (34) compared the application of the levels of evidence synthesis method in different reviews of the same research group. They concluded that there were "markedly different conclusions on treatment efficacy" and they cautioned against its use. Also advocates of the levels of evidence method concluded that the system is sensitive to how the method is interpreted and used (35). However, the main argument against its use is that non-significant results are counted as evidence of no effect even in cases where the confidence intervals are wide and, thus, these studies do not add to the power of the systematic review. This

[1] Consistent findings among multiple high-quality randomized trials add up to strong evidence, multiple low-quality randomized trials or only one high-quality trial result in moderate evidence, only one low-quality trial leads to limited evidence. Conflicting evidence is the outcome when there are inconsistent findings.

increases the chance of a false-negative result or beta-error of not concluding that an intervention is effective even though in reality it is. More generally, the absence of evidence of an effect of an intervention should not be confused with evidence of the absence of an effect (36, 37). The Cochrane Back Review Group has withdrawn the guidance on levels of evidence in its most recent updated guidelines (38).

Best evidence synthesis. Another but similar approach is called best evidence synthesis, which can accommodate studies from a range of disciplines relevant to human health (39). This approach does not differ from the levels of evidence approach described above: study results are synthesized into strong, moderate, partial, or mixed evidence based on the quality and the positive or negative outcome of the study. Slavin, who proposed the method, indeed criticized the prevailing approach of meta-analysis in social sciences at that time, in which study results were combined regardless of methodological quality. He proposed to exclude lower quality evidence in case higher quality evidence is available and thus always base conclusions on the best available evidence. He further proposed to conduct proper meta-analysis of the results of included studies but to finally also comment on and describe more than just the effect-sizes resulting from the meta-analysis.

Worked example

Using an example, we would like to point out that the levels of evidence approach can lead to conclusions that are different from those obtained with a proper meta-analysis. Hartvigsen et al (40) performed a systematic review of the relation between psychosocial factors at work and the presence of back pain (40). The authors used a system of levels of evidence to assess the association between organizational aspects of work and back pain. They included only prospective cohort studies that compared the occurrence of back pain between workers with high and low levels of exposure to psychosocial factors at work. Based on nine studies, they concluded that there was moderate evidence for no association between organizational stress and low-back pain (41–49).

We reanalyzed their material with the procedure described above and combined the results in a meta-analysis.

We took the list of nine included studies as a point of departure and re-analyzed them using the decision flowchart provided in figure 1. Two articles reported on the same study and thus we excluded one [personal communication, Gonge et al (49)]. Most studies reported on more than one measure of organizational stress. We used the job–demand–control model of Karasek (50) to group the exposures according to psychological demands, skill discretion, and decision authority (table 1). In all studies, the control condition had a much lower degree of exposure or no exposure with sufficient contrast to bring about a difference in outcome. Participants were not similar in the studies varying from general population to construction workers, but we assumed that the effects of stress exposure would be similar. We also assumed that effects would not vary according to gender but where effect sizes were reported separately for men and women, those for men were used. The study designs were all prospective cohort studies except for the study by Gonge et al (48) that used a case-crossover design. This design is substantially different from the other studies and so we excluded it. The outcome measures were all self-reports of low-back symptoms that we thought would be similarly influenced by organizational stress. Follow-up times varied from 1–10 years and were all long enough to bring about an effect of organizational stress. Effect sizes were however different across studies. In three studies, back-pain scores were analyzed as continuous variables using multiple regression analysis. Because articles reported only betas and P-values and not standard errors, we could not combine these effect sizes. In four other studies, dichotomous variables were used and analyzed with logistic regression analysis. We combined effect sizes based on dichotomous outcomes using the generic inverse variance method as implemented in Revman 5.1 (Cochrane Collaboration, Copenhagen, Denmark). As input in Revman, we used the natural logarithm of the OR [ln(OR)] and its standard error which we calculated from the 95% confidence intervals (95% CI) provided in the articles. Because statistical heterogeneity was low, we used a fixed-effects model. We followed the same procedure for psychological demands and skill discretion.

For the relation between psychological demands and low-back pain, this resulted in a pooled OR of 1.83 (95% CI 1.25–2.67), which was supported by two studies that used multivariate regression. For decision authority the OR was 1.12 (95% CI 0.73–1.72), which was also supported by two studies that used multivariate regression. For skill discretion, the OR was 1.27 (95% CI 0.89–1.80) supported by two studies that used multivariate regression (figure 2). In contrast to Hartvigsen et al's conclusions, based on these new results, we found evidence that psychological demands at work are related to low-back pain and that there is a possible but uncertain relation between decision authority and skill discretion and low-back pain. This change in conclusion is mainly due to the use of meta-analysis instead of the levels of evidence approach. Better classification of the exposure categories and stricter application of the inclusion criteria did not change the results.

Table 1. Characteristics of studies included in worked example of meta-analysis of psychosocial factors and back pain from Hartvigsen et al (40). [LBP=low-back pain; NS=not significant; OR=odds ratio; PR=prevalence ratio; RR=rate ratio; 95% CI=95% confidence interval]

Authors	Study design	Participants	Follow-up time (years)	Partial correlation	RR	OR	Beta	PR	95% CI	P-value	Outcome
Leino et al (41)	Prospective cohort	Male blue-collar factory workers N=149	10								Change in Back pain during last 12 months score over 10 years (1–4)
Psychological demands: overstrain				0.05						0.05	
Skill discretion: work content				0.07						>0.01	
Decision authority: work control				0.05						>0.05	
Bildt et al (42)	Prospective cohort	Various occupations N=420	4								Incident LBP
Psychological demands: job strain						2.2			0.8–5.8		
Skills discretion: job development						1.4			0.7–3.0		
Hoogendoorn et al (43)	Prospective cohort	Various occupations; 30% female N=861	3								Incident LBP
Psychological demands: quantitative demand					1.41				0.56–1.71		
Skills discretion: skills discretion					0.97				0.53 –1.75		
Decision authority: decision authority					0.98				0.56– 1.71		
Shannon et al (44)	Prospective cohort	Hospital workers; 12% male N=350	1								Change in back pain score
Psychological demands: job demands							0.12			0.03	
Skills discretion: job influence							-0.12			0.02	
Decision authority: decision latitude							NS				
Torp et al (45)	Prospective cohort	Garage workers; 2% female N=630	1								Back pain in last 30 days
Psychological demands: psychological demands							0.005			>0.05	
Decision authority: decision authority							- 0.078			<0.05	
Elfering et al (46)	Prospective cohort	Female nurses N=114	1								Back pain more than once per month in the past year
Psychological demands: time control						4.61			1.42–15.03		
Latza et al (47)	Prospective cohort	Male construction workers N=488	3								Back pain more than 90 days in past year
Psychological demands: time pressure								1.7	0.92–3.15		
Skills discretion: monotonous work								1.5	0.86–2.62		
Decision authority: job control								1.39	0.69–2.83		

Concluding remarks

Synthesis of studies in systematic reviews asks especially for judgment on the conceptual similarity of studies. Such a judgment will lead more often to proper meta-analysis or narrative synthesis. Alternatives such as vote counting, levels of evidence synthesis, or best evidence synthesis are better avoided because they may produce biased results of systematic reviews.

References

1. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ. 1996 Jan 13;312(7023):71–2.
2. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 2009 Jul 21;6(7):e1000097. http://dx.doi.org/10.1371/journal.pmed.1000097.
3. Schonstein E, Verbeek JH. Occupational health systematic reviews: An overview. Work. 2006;26(3):255–8.
4. Verbeek J. More systematic reviews needed to improve

occupational health. Scand J Work Environ Health. 2007 Apr;33(2):81–3.

5. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci. 2010;5:69. http://dx.doi.org/10.1186/1748-5908-5-69.

6. Dixon-Woods M, Agarwal S, Jones D, Young B, Sutton A. Synthesising qualitative and quantitative evidence: a review of possible methods. J Health Serv Res Policy. 2005 Jan;10(1):45–53. http://dx.doi.org/10.1258/1355819052801804.

7. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to meta-analysis. Chichester, UK: John Wiley and Sons; 2010

8. Cooper HM, Hedges LV. The handbook of research-synthesis. New York: Russell Sage Foundation; 1994.

9. Egger M, Davey Smith G, O'Rourke K. Rationale, potentials and promise of systematic reviews. In: Egger M, Davey Smith G, Altman D, editors. Systematic Reviews in Health Care; meta-analysis in context. 2nd ed. London: BMJ Publishing Group; 2001. p3–22.

10. Palmer KT, Harris EC, Linaker C, Barker M, Lawrence W, Cooper C, et al. Effectiveness of community- and workplace-based interventions to manage musculoskeletal-related sickness absence and job loss-a systematic review. Rheumatology (Oxford). 2011 Mar 16. [Epub ahead of print]

11. Henken HT, Huibers MJ, Churchill R, Restifo K, Roelofs J. Family therapy for depression. Cochrane Database Syst Rev. 2007;(3):CD006728.

12. Forster A, Lambley R, Hardy J, Young J, Smith J, Green J, et al. Rehabilitation for older people in long-term care. Cochrane Database Syst Rev. 2009;(1):CD004294.

13. Ioannidis JP, Patsopoulos NA, Rothstein HR. Reasons or excuses for avoiding meta-analysis in forest plots. BMJ. 2008 Jun 21;336(7658):1413–5. http://dx.doi.org/10.1136/bmj.a117.

14. Krebs P, Prochaska JO, Rossi JS. A meta-analysis of computer-tailored interventions for health behavior change. Prev Med. 2010 Sep;51(3-4):214–21. http://dx.doi.org/10.1016/j.ypmed.2010.06.004.

15. Hillsdon M, Foster C, Thorogood M. Interventions for promoting physical activity. Cochrane Database Syst Rev. 2005;(1):CD003180.

16. Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. BMJ. 2011;342:d549. http://dx.doi.org/10.1136/bmj.d549.

17. Schaafsma F, Schonstein E, Ojajarvi A, Verbeek J. Physical conditioning programs for improving work outcomes among workers with back pain. Scand J Work Environ Health. 2011 Jan;37(1):1–5. http://dx.doi.org/10.5271/sjweh.3078.

18. Glasziou P, Meats E, Heneghan C, Shepperd S. What is missing from descriptions of treatment in trials and reviews? BMJ. 2008 Jun 28;336(7659):1472–4. http://dx.doi.org/10.1136/bmj.39590.732037.47.

19. Freak-Poli R, Cumpston M, Peeters A, Clemes SA. Workplace pedometer interventions for increasing physical activity (protocol). Cochrane Database Syst Rev. Issue 7, CD009209. 2011.

20. Kok G, Mesters I. Getting inside the black box of health promotion programmes using Intervention Mapping. Chronic Illn. 2011 Sep;7(3):176–80. http://dx.doi.org/10.1177/1742395311403013.

21. Sauni R, Uitti J, Jauhiainen M, Kreiss K, Sigsgaard T, Verbeek J. Remediating buildings damaged by dampness and mould for preventing respiratory tract symptoms, infections and asthma. Cochrane Database Syst Rev. 2009;(3):CD007897.

22. Verbeek J, Martimo KP, Karppinen J, Takala EP, Kuijer PP, et al. Effect of training and lifting equipment for preventing back pain in lifting and handling: systematic review. Cochrane Database Syst Rev. 2011 Feb 23.

23. Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions. Version 5.0.2. 2011. The Cochrane Collaboration.

24. de Groene GJ, Pal TM, Beach J, Tarlo SM, Spreeuwers D, Frings-Dresen MH, et al. Workplace interventions for treatment of occupational asthma. Cochrane Database Syst Rev. 2011;5:CD006308.

25. Furlan AD, Tomlinson G, Jadad AA, Bombardier C. Examining heterogeneity in meta-analysis: comparing results of randomized trials and nonrandomized studies of interventions for low back pain. Spine (Phila Pa 1976). 2008 Feb 1;33(3):339–48.

26. Anzures-Cabrera J, Higgins JP. Graphical displays for meta-analysis: An overview with suggestions for practice. Research Synthesis Methods 2010;1:66–80. http://dx.doi.org/10.1002/jrsm.6.

27. Glasziou PP, Sanders SL. Investigating causes of heterogeneity in systematic reviews. Stat Med. 2002 Jun 15;21(11):1503–11. http://dx.doi.org/10.1002/sim.1183.

28. Higgins JP, Thompson SG. Controlling the risk of spurious findings from meta-regression. Stat Med. 2004 Jun 15;23(11):1663–82. http://dx.doi.org/10.1002/sim.1752.

29. Popay J, Baldwin S, Arai L, Britten N, Petticrew M, Rodgers M, et al. Narrative synthesis in systematic reviews. 2006. Report No: Methods Briefing 22.

30. Rodgers M, Sowden A, Petticrew M, Arai L, Robers H, Britten N, et al. Testing Methodological Guidance on the Conduct of Narrative Synthesis in Systematic Reviews, effectiveness of Interventions to Promote Smoke Alarm Ownership and Function. Evaluation. 2009;15(1):49–74. http://dx.doi.org/10.1177/1356389008097871.

31. Sutedja NA, Fischer K, Veldink JH, van der Heijden GJ, Kromhout H, Heederik D, et al. What we truly know about occupation as a risk factor for ALS: a critical and systematic review. Amyotroph Lateral Scler. 2009 Oct;10(5-6):295–301. http://dx.doi.org/10.3109/17482960802430799.

32. van Tulder M, Furlan A, Bombardier C, Bouter L. Updated method guidelines for systematic reviews in the cochrane collaboration back review group. Spine (Phila Pa 1976). 2003 Jun 15;28(12):1290–9.

33. Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schunemann HJ. What is "quality of evidence" and why is it important to clinicians? BMJ. 2008 May 3;336(7651):995–8. http://dx.doi.org/10.1136/bmj.39490.551019.BE.

34. Ferreira PH, Ferreira ML, Maher CG, Refshauge K, Herbert RD, Latimer J. Effect of applying different "levels of evidence" criteria on conclusions of Cochrane reviews of interventions for low back pain. J Clin Epidemiol. 2002 Nov;55(11):1126–9. http://dx.doi.org/10.1016/S0895-4356(02)00498-5.

35. van der Velde G, vanTulder M, Cote P, Hogg-Johnson S, Aker P, Cassidy JD, et al. The sensitivity of review results to methods used to appraise and incorporate trial quality into data synthesis. Spine (Phila Pa 1976). 2007 Apr 1;32(7):796–806.

36. Altman D, Bland JM. Confidence intervals illuminate absence of evidence. BMJ. 2004 Apr 24;328(7446):1016–7. http://dx.doi.org/10.1136/bmj.328.7446.1016-b.

37. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ. 1995 Aug 19;311(7003):485.

38. Furlan AD, Pennick V, Bombardier C, van TM. 2009 updated method guidelines for systematic reviews in the Cochrane Back Review Group. Spine (Phila Pa 1976). 2009 Aug 15;34(18):1929–41.

39. Rivilis I, Van ED, Cullen K, Cole DC, Irvin E, Tyson J, et al. Effectiveness of participatory ergonomic interventions on health outcomes: a systematic review. Appl Ergon. 2008 May;39(3):342–58. http://dx.doi.org/10.1016/j.apergo.2007.08.006.

40. Hartvigsen J, Lings S, Leboeuf-Yde C, Bakketeig L. Psychosocial factors at work in relation to low back pain and consequences of low back pain; a systematic, critical review of prospective cohort studies. Occup Environ Med. 2004 Jan;61(1):e2.

41. Leino PI, Hanninen V. Psychosocial factors at work in relation to back and limb disorders. Scand J Work Environ Health. 1995 Apr;21(2):134–42.

42. Bildt C, Alfredsson L, Michelsen H, Punnett L, Vingård E, Torgen M, et al. Occupational and nonoccupational risk indicators for incident and chronic low back pain in a sample of the Swedish general population during a 4-year period: an influence of depression? Int J Behav Med. 2000;7(4):372–92. http://dx.doi.org/10.1207/S15327558IJBM0704_07.

43. Hoogendoorn WE, Bongers PM, de Vet HC, Houtman IL, Ariens GA, van MW, et al. Psychosocial work characteristics and psychological strain in relation to low-back pain. Scand J Work Environ Health. 2001 Aug;27(4):258–67.

44. Shannon HS, Woodward CA, Cunningham CE, McIntosh J, Lendrum B, Brown J, et al. Changes in general health and musculoskeletal outcomes in the workforce of a hospital undergoing rapid change: a longitudinal study. J Occup Health Psychol. 2001 Jan;6(1):3–14. http://dx.doi.org/10.1037/1076-8998.6.1.3.

45. Torp S, Riise T, Moen BE. The impact of psychosocial work factors on musculoskeletal pain: a prospective study. J Occup Environ Med. 2001 Feb;43(2):120–6. http://dx.doi.org/10.1097/00043764-200102000-00010.

46. Elfering A, Grebner S, Semmer NK, Gerber H. Time control, catecholamines and back pain among young nurses. Scand J Work Environ Health. 2002 Dec;28(6):386–93.

47. Latza U, Pfahlberg A, Gefeller O. Impact of repetitive manual materials handling and psychosocial work factors on the future prevalence of chronic low-back pain among construction workers. Scand J Work Environ Health. 2002 Oct;28(5):314–23.

48. Gonge H, Jensen LD, Bonde JP. Do psychosocial strain and physical exertion predict onset of low-back pain among nursing aides? Scand J Work Environ Health. 2001 Dec;27(6):388–94.

49. Gonge H, Jensen LD, Bonde JP. Are psychosocial factors associated with low-back pain among nursing personnel? Work & Stress. 2002;16(1):79–87. http://dx.doi.org/10.1080/02678370110111985.

50. Karasek R, Choi B, Ostergren PO, Ferrario M, De SP. Testing two methods to create comparable scale scores between the Job Content Questionnaire (JCQ) and JCQ-like questionnaires in the European JACE Study. Int J Behav Med. 2007;14(4):189–201. http://dx.doi.org/10.1007/BF03002993.

Received for publication: 1 July 2011