Author's response to reviews

Title:A systematic review including meta analysis of work environment and depressive symptoms Running title: Job strain and bullying at work are related to depressive symptoms

Authors:

Töres Theorell (Tores.Theorell@stressforskning.su.se)
Anne Hammarström (anne.hammarstrom@fammed.umu.se)
Gunnar Aronsson (Gunnar.Aronsson@psychology.su.se)
Lil Träskman Bendz (lil.traskman_bendz@med.lu.se)
Tom Grape (tom.grape@tiohundra.se)
Christer Hogstedt (christer.hogstedt@gmail.com)
Ina Marteinsdottir (ina.marteinsdottir@liu.se)
Ingmar Skoog (ingmar.skoog@neuro.gu.se)
Charlotte Hall (Charlotte.Hall@sbu.se)

Version:3 Date:20 May 2015

Author's response to reviews: see over
Reviewer's report (1)

Title: A systematic review including meta analysis of work environment and depressive symptoms Running title: Job strain and bullying at work are related to depressive symptoms

Version: 2 Date: 23 March 2015 Reviewer: Jian Li Wang Reviewer's report:

This is a systematic review on relationship between workplace factors and depression, which is an important research topic. This review focused on depression only and a broad spectrum of workplace factors. The conclusions are consistent with the literature that high job strain and bullying have significant impacts on the development of depression. I offer several comments on the paper for the authors’ consideration.

Major compulsory revisions:

First, the review included studies that used structured diagnostic interviews or depression rating scales. The outcomes based on these instruments could be either major depressive episode/disorder based on the DSM or ICD criteria, or “depression” that exceeds certain cut-offs. Whereas in the manuscript, the authors used the term of “depressive symptoms” which is a very broad concept. “Depressive symptoms” may include one or two symptoms, which can be normal mood changes; it may also refer several symptoms that exceed the pre-defined threshold. To be distinguished from normal mood changes, why not use the term “depression” or “severe depressive symptoms”, or other appropriate terms?

We have added the following paragraph to the Method section in order to clarify this important topic. In addition, the Discussion includes information about homogeneity tests which showed that the findings with self-reported symptom scores were similar to those with standardized interview based depression.

The following was added:

“3, The study analyzed symptoms of depression which had been certified through diagnostic investigation or with established scales. We argued that not only diagnosed major depression, but also milder states with depressive symptoms are relevant since depressive feelings give rise to suffering, increase the risk of long term sick leave and cause productivity decline and quality loss in workplaces [10]. Thus, our review included both studies with structured clinical interviews about diagnosed depression and studies with rating scales about depressive symptoms. As diagnosed depression is also to a large extent based on symptoms we decided that the most accurate naming of the outcome of our review was depressive symptoms. “

We had exhaustive discussions about this in the expert group. On one hand the psychiatrists made the obvious argument that clinically defined depression is the only outcome that clinical psychiatrists in our potential readership would take seriously. On the other hand the epidemiologists argued that the societal impact of depressive symptoms is quite substantial – an
argument that we have tried to bring forth in our text when we mention productivity loss etc. Since the published articles are dominated by outcomes defined by means of standardized questionnaires, we decided to make this our main outcome. However we also report whenever possible findings on standardized interview-based depressive states as outcome. For a couple of exposures it was possible to do homogeneity tests which showed that the findings with self-reported symptom scores were similar to those with standardized interview based depression.

Second, I believe that the selected studies used different instruments/scales to measure the exposure variables, e.g., job strain, and the studies might have used different cut-offs or definitions for a specific exposure. It will be helpful if the authors could comment how different definitions of exposure may have affected the conclusions.

We have added the following paragraph to the methods section:

Even between studies of specific work environment factors there were differences with regard to operationalization of exposure. Examples are job strain (combination of high psychological demands and low decision latitude) and effort reward imbalance (combination of high effort and poor reward). Since the overall aim of the present study was to grade total evidence, not to assess magnitude of associations, and since it was impossible to re-construct operationalizations in such a way that they would match one another we decided to use the definitions presented by the authors themselves.

In the discussion section we added the following:

It should be cautioned that the operationalization of job strain differed between studies. The majority of the published studies used the median split definition (above median for the psychological demands score and below median for the decision latitude score). When exposure to job strain is defined in this way and the remaining participants in the study are defined as unexposed there is relatively little contrast between unexposed and exposed subjects. This may lead to underestimation of the true association.

Third, the authors reported no gender/sex difference. Again, I think that it may be related to how exposure is defined and whether the same or different thresholds were used for men and women. Wang et al.’s research indicated that job strain seemed to affect men and women differently in relation to the risk of major depressive disorder. For example, men had elevated risk of MDD only if they were exposed to extremely high level of job strain; women on the other hand had high risk of MDD even they were exposed to moderate job strain. If the same threshold of job strain for men was used for women, the association between job strain and MDD in women disappeared or diluted. You may find the results and also sex specific associations between work-family conflicts, effort-reward imbalance and MDD in Wang et al. Am J Epidemiol. 2012. 176(1): 52-59.

As a response, we have added the following sentences to the Discussion:

Other studies indicate that work conditions can affect men and women differently in relation to development of major depressive disorder (MDD). For example, a Canadian study showed that
men had elevated risk of MDD only if they were exposed to extremely high level of job strain while women had elevated risk of MDD even when exposed to moderate job strain (14). The study points to the need of contextualizing findings about mental health. It may also illustrate that gender could be more relevant for the relationship between working conditions and major depressive disorder than for the relationship between working conditions and depressive symptoms.

Finally, given that there have been several systematic reviews on this topic, it would be helpful if the authors could elaborate in the Discussion on what this review has added to the literature on top of the previous ones.

As a response to your comment we have added the following paragraph to the Discussion:

The review differs from earlier studies in the field due to its comprehensive and thorough approach. Our review is based on an extremely thorough literature search as well as on a well-described and systematic evaluation of a large number of publications. Thus, it includes all kinds of environmental exposures, physical as well as psychosocial, and it is based upon a systematic approach. This is the first review in which the examination of evidence follows (a slight modification of) GRADE principles. Furthermore it is including more recently published research than previous reviews.

Discretionary revisions.

Several longitudinal studies on this topic were not included in the reviews, for example:


and few others. I will leave it to the authors to determine the eligibility.

Yes there are studies which were not included but all of those which the reader will not find have been sorted out because they were judged to be non-relevant or lacked some aspect that was necessary for medium high or high quality using our specific criteria and framework. This does not necessarily mean that these studies lack rigor or scientific value. Also in several cases when many studies have been published on one and the same cohort during the same time period (doublets or triplets) we were forced to select only one of them in order not to create undue weight to such a study. The full lists can be found in http://www.sbu.se/223E.

Level of interest: An article of outstanding merit and interest in its field

Quality of written English: Acceptable

Statistical review: Yes, and I have assessed the statistics in my report. Declaration of competing interests: I declare that I have no competing interests.

Reviewer's report

Title: A systematic review including meta analysis of work environment and depressive
symptoms Running title: Job strain and bullying at work are related to depressive symptoms

Version:2Date:14 April 2015

Reviewer: Karen Nieuwenhuijsen

Reviewer's report:

This review is important in our field. However, the presentation could be much improved and essential information and data are not presented.

Major Compulsory Revisions

1. Aim of the study: The term ‘weighted’ is used in the aim of this study, but this is not introduced in the background. Weighted could mean a number of things, perhaps they are referring to GRADE?

Yes we are referring to GRADE so we have changed the term from weighted to graded.

2. The methods section is too unstructured, in a systematic review I would expect the steps of the review process to be much less narrative and more structured. Subheadings would help, but

3. Important information (inclusion criteria and quality criteria) should be in the main text.

The whole methods section has been re-organized under subheadings. In addition, the inclusion criteria and quality criteria have been added to the main text.

4. Page 5: what is the place in the review process of the expert group. Now it just states how they were recruited while as a reader you do not know why an expert group is needed or what function it has.

Yes we have tried to describe the expert recruitment and its role in a better way

After that the scientific experts started their examination. Pre-set evaluation forms were used. The experts judged relevance and quality of the studies on the basis of the relevance/quality criteria, their experience as researchers and their knowledge of the field. Accordingly they were recruited among Swedish academic high ranking specialists in fields of relevance for the process, namely psychiatry (three), epidemiology and stress research (three), work psychology (one) and family practice (one). This group was divided into pairs with as widely differing specialty in the pair as possible. In the following process, the articles remaining in the process were randomly assigned to the four pairs (with avoidance of author bias). Concordance in judgments of relevance and
quality was trained. After the training session, each member of the pair did the assessments separately, and then discordances were discussed within the pair. If disagreement remained another pair was asked to make an independent judgment. If that decision was in disagreement with the first group, we made the decision in the whole group.

5. Please be more specific on the criteria for design equivalent to prospective (page 6, line 12).

We added the following formulation:

“By case control studies with “design equivalent to prospective” we are referring to studies with strict definition of cases recruited in a representative way in the same population as the control group. Assessments of exposure should have been made before disease onset.”

6. I don’t see why only studies with > 100 participants are included. A meta analysis would take the smaller n into account.

The reason is that very small studies would add very little to a meta- analysis. In addition our main focus has been more on the GRADE procedure than on the magnitude of the relationships and therefore we avoided studies with small statistical power.

7. The levels for the quality rating should be specified in the main text.
   This has been stated in the new version

8. quality aspect: ‘sampling’: most studies were based on.... that is a result, what was the criterion?
   The sentence has been moved to the Result section.

9. quality aspect: ‘sampling’: How is an insightful discussion of the consequences of dropout an aspect of quality usable for the GRADE method? GRADE is all about whether the results may be biased, a discussion of consequences does not change that.

   We have changed the name of the label from Sampling to "Representativeness of study sample”. Systematic non-response was one topic when we valued the quality of the representativeness of the study. If a manuscript describing a study with sizeable and possible systematic attrition contains a discussion - for instance by analyses of socio-demographic characteristics of drop-outs – about likely consequences of that drop-out the discussion may gain credibility and the study accordingly gains in assessed quality

   The latter part of the paragraph now reads: Representativeness and ways of defining and recruiting the sample as well as attrition in different steps were considered in the quality rating. Statistical considerations and an insightful discussion of possible consequences of a possible systematic drop-out on findings were required in case of marked drop-out problems.
COMMENT TO THE REVIEWER: If a manuscript describing a study with sizeable attrition contains a discussion - for instance by analyses of sociodemographic characteristics of drop-outs – about likely consequences of that drop-out the discussion may gain credibility and the study accordingly gains in assessed quality.

10. Overall: in the discussion of quality aspects many times results of the review are presented. Quality criteria should be formulated without knowledge of the findings, this presentation may give the wrong impression.

We have moved the results in the Quality criteria to the Result section. We are here describing all the aspects of judgments that have influenced quality rating in the original publications.

11. Page 7: line 22 to 32: these are results.

The text has been moved to the Result section.

12. Argumentation for GRADE could best be presented in the background section.

Yes we have moved the argument for GRADE. The remaining text could not be regarded as results. In this particular study gradings and all-over assessments of evidence are our results.

13. The GRADE system should be presented in the methods section to provide transparency for the readers. The results of the GRADE procedure belong in the results section.

We have presented the GRADE system in the methods.

GRADE procedure

An important aspect of the systematic review process was to systematically and transparently assess the scientific evidence. According to the GRADE instructions explicit consideration should be given to each of the GRADE criteria for assessing the quality of evidence (risk of bias/study limitations, directness, consistency of results, precision, publication bias, magnitude of the effect, dose-response gradient, influence of residual plausible confounding and bias “antagonistic bias”) although different terminology may be used. For level 4 (=High), randomized trials are required and there were no such published relevant studies in our search. For observational studies of the kind included in the present review, the highest possible grade is Moderate=3 if there is sufficient reason for an upgrading from the normal level for such studies of 2 (=Low). Level 1 (=Very low) corresponds to evidence based on case reports and case series or on reports downgraded evidence from observational studies.

Results of the grading process are presented on page 9

Table 1 shows the results of the evidence grading process. Three exposures, two harmful (job strain and bullying) and one protective (control/decision latitude) were judged to have moderate evidence (grade 3) while 18 exposures were judged to have low (grade 2) evidence. Ten exposures were judged to have very low evidence (grade 1). Three of the exposures judged to
have very low evidence were related to heavy metals and other chemical exposures.

14: More details on the meta analysis should be provided. 15: What are informal homogeneity tests??

Explanations have now been added

**Meta-analyses/Forest plots**
In the studies results were reported as calculations of association, e.g. expressed as odds ratios, from multiple logistic regression, multivariate correlations or multiple linear regression coefficients. Whenever possible, the results were transformed into multiple logistic regression odds ratios. Forest plots were used for visual interpretation. To assist in illustrating the results, and as a contribution to the overall assessment, these forest plots (meta-analyses) were conducted when in at least two studies the same risk factor was analysed and mathematically comparable data was provided using the Comprehensive Meta-Analysis software package (www.meta-analysis.com/index.php). Since the participants in the various studies might be construed as coming from the same population (workers) or from different populations (i.e. according to each study’s inclusion criteria) we chose to use a fixed effects model. The strength of the scientific evidence, using data from all of the included studies (not just those illustrated in the meta-analyses), was determined by pairs of the authors of this paper and then discussed and confirmed by all authors. Informal homogeneity tests were performed in order to compare results from studies using standardized depression interviews versus self-reported questionnaires, high quality versus medium high quality studies, general population studies versus specific occupational cohorts and men versus women. In these tests, we conducted sub-analyses of the presented findings and compared results between the sub-categories, e.g. if the association between job exposure and depressive symptoms differed according to the instrument used for assessing the symptoms.

16: Results: Overall, a clear overview of the findings is not presented. Which factors, which quality levels? And by presenting forest plots of parts of the results only, the data of the meta analysis of the other factors are not shown. An overview of the findings with the data is needed for readers to assess the findings. Page 10 includes many statements without any reference to the direction and strengths of the associations!

OK we have supplemented the results section.

Figure 1 shows the number of articles that were perused in the different steps. The process also included burnout as outcome. The results of the burnout review will be reported elsewhere. Altogether 20 828 articles were screened in the initial search process, and 488 of those were eligible in the review of depressive symptoms (and 202 for the review of burnout). 324 full text articles with depression as outcome were found not to fill inclusion criteria. Hence, 164 studies remained for relevance assessment. 84 of those were judged as not relevant and hence 80 studies were assessed with regard to quality. 19 were judged to be of high, 40 of moderately high and 21 of low quality. The grading of evidence has been based upon the 59 relevant studies with high/medium high quality. A detailed table showing the full results of the data extraction is
Most studies were based on population samples although studies of samples from companies and occupational groups were also present. Few studies that were judged to be relevant were based upon objective assessments of exposure. Subjective assessments based upon standardized and validated questionnaires (for instance demand/control/support, effort/reward, procedural justice and bullying) were used in most studies. The most widely used established questionnaires rendered high quality ratings. With regard to depression outcome, both standardized interviews (mostly Composite International Diagnostic Interview, CIDI) performed by trained interviewers and different versions of standardized questionnaires (such as Center for Epidemiological Studies- Depression Scale, CES-D, and Hospital Anxiety and Depression Scale, HAD, and Hamilton Depression Scale, HRSD) for depressive symptoms were used.

Table 1 shows the results of the evidence grading process. Three exposures, two harmful (job strain and bullying) and one protective (control/decision latitude) were judged to have moderate evidence (grade 3) while 18 exposures were judged to have low (grade 2) evidence. Ten exposures were judged to have very low evidence (grade 1). Three of the exposures judged to have very low evidence were related to heavy metals and other chemical exposures. The most extensively studied factors were decision latitude (158 251 subjects in 19 studies) and job strain - the combination of high psychological demands and low decision latitude (197 682 subjects in 14 studies). It was possible to compute a weighted odds ratio 1.74 (95% CI 1.54 to 1.96 for studies with odds ratio calculations). A high decision latitude protected statistically against worsening depressive symptoms – with a weighted odds ratio of 0.73 (95% CI 0.68 to 0.77). Bullying had been studied in 15 173 subjects in three studies. One of these studies showed results for men and women separately. Despite the relatively small number of studies, bullying was judged to be related to worsening depressive symptoms with an evidence grade of 3 as the findings were very consistent and the odds ratios were high (the weighted odds ratio being 2.82; 95% CI 2.21 to 3.59).

17. Discussion: highly unstructured. methodological considerations are presented together with other parts of the discussion. And then under the subheading ‘on the methods’ results are presented.

We regret that the Discussion was unstructured. It has been re-organized and we have tried to facilitate reading by introduction of several new headings.

Minor Essential Revisions

1. The numbering of the references should be checked. nr7 (page 3/line 34) refers to nr 8 in the reference list. Also the spelling of the references (I only checked reference 8) contains typing errors.

This has been corrected

Discretionary Revisions
Background: the authors point to the Michie review conclusion on how factors are amenable to change. But the concluding line of that paragraph (line 24) could be more specific. How can the re-organisational exposure of factors be explored? If the authors mean that by knowing the associations, interventions can be focused towards these factors (and that is why the review is needed) I agree. But to what extent the psychosocial exposures can be changed cannot be answered by this review.

We have deleted that sentence.

Level of interest: An article of importance in its field
Quality of written English: Acceptable
Statistical review: No, the manuscript does not need to be seen by a statistician.
Declaration of competing interests: □ I declare that I have no competing interests

For the editor: Yes we have now included the reference to the very recent publication by Niedhammer et al.