PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	The impact of fatigue and insufficient sleep on physician and patient outcomes: A systematic review
AUTHORS	Gates, Michelle; Wingert, Aireen; Featherstone, Robin; Samuels, Charles; Simon, Christopher; Dyson, Michele

VERSION 1 – REVIEW

REVIEWER	Meghna P. Mansukhani Mayo Clinic, Rochester, Minnesota, USA
REVIEW RETURNED	05-Feb-2018

GENERAL COMMENTS	I commend the authors for undertaking this work. I have a few minor comments. 1. Was sleepiness included under the outcome of fatigue? If so, would add to the table of outcomes. If the outcome being measured was ESS in several of the studies included here, would change the terminology in the paper to "sleepiness" or "fatigue and sleepiness." 2. What exactly are the methodological weaknesses and biases? Would recommend adding more specifics to the limitations section- e.g. mostly male physicians, many survey-based studies and/or only subjective measures assessed, many studies with high risk of bias, etc. 3. The individual studies are described but it would be helpful to add 1-2 statements at the end of each section summarizing the literature (and major weaknesses, if any). It would be worth mentioning how some of the outcomes were measured (e.g. melatonin levels, ESS, PSQI, performance vigilance testing etc.) in the main text rather than only in the supplement.

REVIEWER	Ilda Amirian Department og gynecology and obstetrics Zealand University Hospital, Roskilde Sygehusvej 10, 4000 Roskilde
REVIEW RETURNED	02-Mar-2018

GENERAL COMMENTS	Thank you for having me review this very interesting systematic review. I find the subject highly relevant, and the review very well performed. Just one comment: in the abstract under strengths and limitations, the authors should avoid superfluous words as "rigourously conducted and transparently reported", it's better to stay neutral. Well done!

REVIEWER	Gordon S. Doig PhD (Epi and Biostats) University of Sydney, Australia
REVIEW RETURNED	18-Mar-2018

GENERAL COMMENTS	Statistical review of manuscript bmjopen-2018-021967 titled “The

	impact of fatigue and sleep deprivation on physician and patient outcomes: A systematic review" 1. Abstract. This reports to be a narrative review, however I have been specifically asked to review statistical methods. By definition, narrative reviews do not contain statistical summaries. If you have analytically synthesized any evidence, please alter your Abstract Methods to report this is a meta-analysis. If you report the results of any analytic summaries, or base any conclusions on these summaries, please report the results in your Abstract Results. 2. Page 7, evidence synthesis. It appears you attempted to pool study results but did not pool due to heterogeneity. In your Methods, report how heterogeneity was assessed and what attempts were made to pool. In your Results, report how you assessed inability to pool. Table 1 suggests you have a number of studies that looked at similar outcomes (Ex. Physician physical and mental health, patient outcomes etc). I do not understand how none of them were able to be pooled. 3. Abstract, conclusions and remainder of Manuscript: If only observational studies demonstrate an association between fatigue and sleep deprivation with physician health, you cannot make the statement that relates causation (Ex. 'impact on' or 'effects' or 'predicted by'). You must use 'associated with'. 4. Example of potential to pool: "One small (n = 11) before-after study showed longer reaction times (690.8• }73.4 vs. 746.5• }113.7 milliseconds) and reduced concentration ability (26.4• }23.5 vs. 56.3• }23.0 on a 100-point scale, P = 0.007) following a 24-hour shift with sleep deprivation[45]; Two others found that sleep loss was associated with slower reaction times.[38, 54]." Why can this data not be pooled? 5. Did this review serve as an Introduction to a Thesis? It is overly broad and would benefit from an increase in focus. At the very least, because the authors appear to have attempted to pool and meta-

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1, Meghna P. Mansukhani

I commend the authors for undertaking this work. I have a few minor comments.

Thank you.

1. Was sleepiness included under the outcome of fatigue? If so, would add to the table of outcomes. If the outcome being measured was ESS in several of the studies included here, would change the terminology in the paper to "sleepiness" or "fatigue and sleepiness."

Thank you for identifying this. We did include sleepiness as an exposure in this review, which we have articulated in the first paragraph of our inclusion criteria: “Exposures of interest included fatigue, sleep restriction, or sleepiness.” We used the terminology presented by authors of the individual studies within the review, and considered sleepiness as a sleep-related exposure. In table 1, we have changed the exposures to read as ‘fatigue-related’ and ‘sleep-related’ to avoid any confusion arising from heterogeneity in how the exposure was measured across studies.

We note that of 37 studies reporting on sleep-related exposures, only 6 (16%) reported on sleepiness, and 5 used the Epworth Sleepiness Scale. Other studies reported sleep hours, sleep deprivation, sleep quality and sleep disruption. Thus, it does not seem accurate to change the terminology to ‘sleepiness’. Instead, we updated the review to (a) use the terms ‘fatigue’ and ‘sleep restriction’, and (b) define what these terms encompass within the results. We thereafter used the term ‘sleep restriction’ when speaking generally about the evidence, and the terminology specific to each study, when relevant. The third paragraph under ‘included study characteristics’ has been modified as follows:

“Fifteen (32%) studies reported on fatigue-related exposures (e.g., as a source of stress, exhaustion, physical fatigue; hereafter referred to as ‘fatigue’), while others (n = 37/47, 79%) reported on sleep-related exposures (e.g., sleep hours, sleep restriction, sleep deprivation, sleep disruption, sleepiness; hereafter referred to as ‘sleep restriction’).”

We have made minor changes throughout the review to adhere to the terminology described above.

2. What exactly are the methodological weaknesses and biases? Would recommend adding more specifics to the limitations section- e.g. mostly male physicians, many survey-based studies and/or only subjective measures assessed, many studies with high risk of bias, etc.

Thank you for identifying this point for clarification. The biases identified are available within Supplement 3 and vary based on design. We have tried to make this more evident within the results (risk of bias appraisal): “Detailed assessments of the sources of bias within each study are shown in Supplementary file 3.”

Sources of bias included the relative lack of control groups, use of samples that were not necessarily representative of the population, subjective measurement of exposure and outcomes, and low response rates for surveys. Other weaknesses included the heavy reliance on cross-sectional designs, small sample sizes, and inclusion of predominantly male physicians within urban settings. We have added some detail to the limitations, as follows:

“While we have identified a diverse body of evidence, we could not draw definitive conclusions due to methodological weaknesses (e.g., 62% at high risk of bias, reliance primarily on cross-sectional designs and uncontrolled studies, subjective measurement of exposures and outcomes, small sample

sizes, inclusion of predominantly male physicians within urban settings) and heterogeneous outcome measures in the included studies."

3. The individual studies are described but it would be helpful to add 1-2 statements at the end of each section summarizing the literature (and major weaknesses, if any). It would be worth mentioning how some of the outcomes were measured (e.g. melatonin levels, ESS, PSQI, performance vigilance testing etc.) in the main text rather than only in the supplement.

Thank you. In the original submission, we attempted to include a broad view of the evidence at the beginning of each results paragraph. Given potential lack of clarity, we have returned to these paragraphs and made amendments to improve the comprehensibility of the evidence summaries, and added information on quality of the included studies. Where possible, we have also added information on the measurement tools used. A sample paragraph for burnout is shown below. Please refer to the main document to view the changes made to each paragraph of the results.

"Seven cross-sectional studies reported on burnout (5 low, 1 unclear, 1 high risk of bias) among surgeons, anesthesiologists, generalists and other mixed groups. Two studies reported on surgeons; the larger (n = 2,564, low risk of bias) study of neurosurgeons showed increased odds of burnout with sleep deprivation (hours of sleep per night; OR 0.84, 95% CI 0.75-0.94, P = 0.002). Among anesthesiologists one study (n = 565, low risk of bias) indicated that burnout (measured via Maslach Burnout Inventory) was more prevalent among the sleep-deprived ('lack of sleep' on one question; 47.6% vs. 16.3%, P < 0.001). In one small (n = 11) study of generalists, those with burnout (measured via Pines Burnout Measure) had poorer Pittsburgh Sleep Quality Index scores (7.24±4.17 vs. 2.72±2.22, P < 0.001). In the two larger studies of mixed physician groups (low risk of bias), burnout (measured via 5-point scale) was more prevalent among those who were sleep deprived (<7 hours of sleep per 24 hours; 39.6% vs. 26.4%, P < 0.05), and physical fatigue ('feeling tired' on a 7-point scale) was correlated with burnout (Shirom-Melamed Burnout Measure; r = 0.88, P < 0.05). In summary, evidence from 7 cross-sectional studies (71% at low risk of bias), showed associations between sleep restriction and burnout."

Reviewer: 2, Ilda Amirian

Thank you for having me review this very interesting systematic review. I find the subject highly relevant, and the review very well performed. Just one comment: in the abstract under strengths and limitations, the authors should avoid superfluous words as "rigourously conducted and transparently reported", it's better to stay neutral. Well done!

Thank you for the kind comments. We have updated the strengths and limitations based on your suggestion and that of the editor, and the tone is now objective:

"The review was informed by the methods outlined by Cochrane and is reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines."

Reviewer: 3, Gordon S. Doig PhD (Epi and Biostats)

1. Abstract. This reports to be a narrative review, however I have been specifically asked to review statistical methods. By definition, narrative reviews do not contain statistical summaries. If you have analytically synthesized any evidence, please alter your Abstract Methods to report this is a meta-analysis. If you report the results of any analytic summaries, or base any conclusions on these summaries, please report the results in your Abstract Results.

Thank you for this guidance. Indeed, the original manuscript reports on a systematic review, not a meta-analysis, which is the reason that terms related to meta-analysis were left out of the title, abstract, methods and results. Based on the comments that follow, we have performed meta-analysis for select outcomes (operating time, operative complications, patient mortality and length of stay).

We have revised the last line of the 'design' heading in our abstract to read: "We pooled findings via meta-analysis when appropriate, or narratively."

We have also revised the results of the abstract as follows: "Of 16,154 records identified, we included 47 quantitative studies of variable quality. 28 studies showed associations between fatigue or sleep restriction and physician health and well-being outcomes. 21 studies showed no association with surgical performance, and mixed findings for psychomotor performance, work performance, and medical errors. We pooled data from six cohort studies for patient outcomes. For sleep deprived versus non-sleep deprived surgeons, we found no difference in patient mortality (n = 60,436, RR 0.98, 95% CI 0.84 to 1.15, p = 0.82, I^2 = 0%), intraoperative complications (n = 19,798, RR 1.35, 95% CI 0.82 to 2.21), postoperative complications (n = 60,201, RR 0.99, 95% CI 0.95 to 1.03) or length of stay (n = 50,046, MD -0.33, 95% CI -1.03 to 0.36)."

2. Page 7, evidence synthesis. It appears you attempted to pool study results but did not pool due to heterogeneity. In your Methods, report how heterogeneity was assessed and what attempts were made to pool. In your Results, report how you assessed inability to pool. Table 1 suggests you have a number of studies that looked at similar outcomes (Ex. Physician physical and mental health, patient outcomes etc). I do not understand how none of them were able to be pooled.

We considered clinical and methodological heterogeneity in our decision to proceed (or not) with meta-analysis. In most cases, even within a category of outcome, the specific outcomes measured within individual studies were highly variable (e.g., mental health might include depression, anxiety, overall mood state, among others). The disparate way in which fatigue and sleep restriction have been described in the literature also contributed to heterogeneity. Examples of exposures included physical or mental fatigue, sleep hours, sleep deprivation, sleep quality, insomnia, and overnight shifts, measured by various validated and non-validated scales. It was rare that the same outcome measure was used more than once among the included studies. Additionally, our review was not restricted by design, and we did not deem it appropriate to combine varied study designs within one meta-analysis. However, based on your suggestion we returned to the data and undertook meta-analysis for outcomes where it was deemed appropriate: operating time, patient mortality, and length of hospital stay. Therefore, we have updated the 'evidence synthesis' section of our methods as follows:

"We considered clinical and methodological heterogeneity in our decision on whether to proceed with meta-analysis for the outcomes identified. For most outcomes, we found insufficient homogeneity in study design, populations, exposures or interventions, and outcome measures to pool the data via meta-analysis. Thus, we have presented the findings for most outcomes narratively and in summary tables.

When statistical pooling was appropriate, this was undertaken using Review Manager (RevMan v.5.3, Copenhagen: The Nordic Cochrane Centre, the Cochrane Collaboration, 2014) via pairwise meta-analysis using the DerSimonian and Laird random effects model (given expected heterogeneity). We summarized dichotomous outcomes using the relative risk (95% confidence interval (CI)) and continuous outcomes using the mean difference (95% CI) since the units across studies were consistent (i.e., minutes). When meta-analysis was conducted, we assessed statistical heterogeneity using the chi-square test (using P = 0.05 as the threshold for significance), and quantified the extent of heterogeneity using the I^2 statistic. We intended to assess small study bias visually by inspecting funnel plots and statistically using Egger's regression test, but did not due to the small number of studies (i.e., <8 per outcome) included in the meta-analyses."

We have updated the results section to include a report of these analyses:

"We pooled the data from these studies[31, 32, 41, 63] via meta-analysis, which showed no difference in operating time (sometimes referred to as surgeon efficiency) between sleep deprived and non-sleep deprived surgeons (Figure 2; n = 50,046, MD -0.14, 95% CI -1.60 to 1.33, P = 0.86, I^2 = 0%)."

and

"We pooled data (collected by chart review) via meta-analysis for procedures performed sleep deprived vs. non-sleep deprived surgeons (or obstetrician-gynecologists in one case). Analyses showed no difference in the rate of intra-operative complications (Figure 3, 3 studies, n = 19,798, RR 1.35, 95% CI 0.82 to 2.21, p=0.24, I^2 = 82%), post-operative complications (Figure 4; 5 studies, n = 60,201, RR 0.99, 95% CI 0.95 to 1.03, p = 0.51, I^2 = 0%), patient mortality (Figure 5; 5 studies, n = 60,436, RR 0.98, 95% CI 0.84 to 1.15, p = 0.82, I^2 = 0%), or length of hospital stay in days (Figure 6; 4 studies, n = 50,046, MD -0.33, 95% CI -1.03 to 0.36, p = 0.35, I^2 = 86%). One study in the mortality analysis reported the number of deaths only as ≤5. We assumed 2 events for this study (midpoint between 0 and 5); sensitivity analysis using the lowest (i.e., 0) and highest (i.e., 5) possible number of events did not change the result (Supplementary file 5). We imputed the average variance for one study[1] in the length of stay analysis; sensitivity analysis using either the highest or lowest SD did not change the results (Supplementary file 5). Subgroup analysis by type of surgery did not explain the substantial between-study heterogeneity detected for length of stay, nor intraoperative complications, though it may be noted that the types of complications reported varied by study."

3. Abstract, conclusions and remainder of Manuscript: If only observational studies

demonstrate an association between fatigue and sleep deprivation with physician health, you cannot make the statement that relates causation (Ex. 'impact on' or 'effects' or 'predicted by'). You must use 'associated with'.

Thank you for noting this. Within the abstract, and throughout the manuscript thereafter, we have removed causal inferences where only observational studies contributed evidence, and instead have reported associations. Please see minor changes in wording throughout the manuscript.

4. Example of potential to pool: "One small (n = 11) before-after study showed longer reaction times (690.8 }73.4 vs. 746.5 }113.7 milliseconds) and reduced concentration ability (26.4 }23.5 vs. 56.3 }23.0 on a 100-point scale, P = 0.007) following a 24-hour shift with sleep deprivation [45]; Two others found that sleep loss was associated with slower reaction times. [38, 54]." Why can this data not be pooled?

Thank you for this example, and please refer to our response to comment #1. In this case, we deemed the studies too heterogeneous to pool. The studies differed in design (one uncontrolled before-after, two cross-sectional), the measures of reaction time used (psychomotor vigilance performance task or test described by Kerr et al. (1992)), and circumstances of the testing. In the study by Lederer (2006), reaction time was measured before and after an on-call shift, whereas Gander (2008) measured reaction time after a series of variable shifts. Finally, Saadat (2017) measured reaction time in the same physicians after regular working hours vs. overnight on-call duty (timing of assessments are described in Supplementary File 5). We did not feel that it would be appropriate to combine these studies, nor that pooling the data would enhance the results as presented.

Based on your suggestion, we have revisited the data and presented pooled effect estimates for the following outcomes: surgeon efficiency; patient outcomes of mortality, complications, length of hospital stay. This is detailed in our responses to comments #1 and #2.

5. Did this review serve as an Introduction to a Thesis? It is overly broad and would benefit from an increase in focus. At the very least, because the authors appear to have attempted to pool and meta-analyze data, explicit evidence of inability to pool must be presented within the primary manuscript.

Thank you for this comment. This review was not the introduction to a thesis; please see our response to the Associate Editor regarding the scope of the review on page 1 of this document. Please see our responses to your previous suggestions regarding pooling the data statistically.

FORMATTING AMENDMENTS (if any)

Kindly re-upload SUPPLEMENTARY FILES in PDF format.

The supplementary files have been re-uploaded as PDFs.

1. Chu MW, Stitt LW, Fox SA, Kiaii B, Quantz M, Guo L, et al. Prospective evaluation of consultant surgeon sleep deprivation and outcomes in more than 4000 consecutive cardiac surgical procedures. Arch Surg. 2011;146(9):1080-5.

VERSION 2 – REVIEW

REVIEWER	Meghna Mansukhani, MD Mayo Clinic, Rochester, MN, USA
REVIEW RETURNED	07-May-2018

GENERAL COMMENTS	Generally, "sleep restriction" is a term used in research where subjects' sleep is experimentally curtailed. I know the authors defined how they used this term in the manuscript but perhaps it might be better to use "insufficient sleep" I leave it up to the editor's discretion.

REVIEWER	Gordon Doig University of Sydney, Australia.
REVIEW RETURNED	21-May-2018

GENERAL COMMENTS	Review of manuscript bmjopen-2018-021967.R1 titled “The impact of fatigue and sleep restriction on physician and patient outcomes: A systematic review” 1. Each time you report results of a meta-analysis in the abstract, manuscript or figures, please report I2 and p-value for heterogeneity immediately after the overall p-value for effect. 2. The presence of important ‘statistical’ heterogeneity infers the presence of ‘clinical’ or ‘methodological’ heterogeneity. Please minimise subjective author decisions with regards to pooling. In your manuscript, please provide a working definition of excessive statistical heterogeneity based on I2 and p-value and do not pool when these thresholds are exceeded. Because of the low-power of the statistical test for heterogeneity, a threshold of 0.10 is usually used for the p-value of heterogeneity. In addition, because of the low power of the statistical test, ‘excessive’ heterogeneity is also defined based on a threshold value of I2, regardless of what the p-value for heterogeneity was. On Page 9, many of your pooled results demonstrate extreme (I2 > 70%) heterogeneity making interpretation potentially meaningless. Use of a random effects model does not make interpretation of analysis with extreme heterogeneity meaningful. 3. Please do not switch between the term ‘heterogeneity’ and ‘homogeneity’ to refer to the same phenomena. Failure to find statistical heterogeneity does not infer homogeneity. It just infers that you did not find statistical heterogeneity. 4. This review remains excessively broad. A focus on either patient outcomes or clinician outcomes would improve this review, along

	with a pre-defined primary outcome. As it stands, I suggest all results should be discounted for multiple-comparisons.

VERSION 2 – AUTHOR RESPONSE

Authors' Response to Reviewer Comments

Reviewer 1: Meghna Mansukhani, MD

Generally, "sleep restriction" is a term used in research where subjects' sleep is experimentally curtailed. I know the authors defined how they used this term in the manuscript but perhaps it might be better to use "insufficient sleep" I leave it up to the editor's discretion.

Thank you. We agree that this change in terminology would be reasonable and have replaced all instances of 'sleep restriction' in the manuscript with 'insufficient sleep'.

Reviewer 3: Gordon Doig

1. Each time you report results of a meta-analysis in the abstract, manuscript or figures, please report I2 and p-value for heterogeneity immediately after the overall p-value for effect.

Thank you. It has been argued that some degree of statistical heterogeneity is inevitable in meta-analysis, which limits the relevance of presenting p-values for the chi-squared statistic.[1] For this reason, we have not added the p-values of the chi-square statistic to our findings. We have nevertheless made these available for interested readers within supplementary file 5. These are also shown on the individual forest plots. We have carefully reviewed the revised manuscript to ensure that the I^2 value is always presented along with the results of meta-analyses.

2. The presence of important 'statistical' heterogeneity infers the presence of 'clinical' or 'methodological' heterogeneity. Please minimise subjective author decisions with regards to pooling. In your manuscript, please provide a working definition of excessive statistical heterogeneity based on I2 and p-value and do not pool when these thresholds are exceeded. Because of the low-power of the statistical test for heterogeneity, a threshold of 0.10 is usually used for the p-value of heterogeneity. In addition, because of the low power of the statistical test, 'excessive' heterogeneity is also defined based on a threshold value of I2, regardless of what the p-value for heterogeneity was. On Page 9, many of your pooled results demonstrate extreme (I2 > 70%) heterogeneity making interpretation potentially meaningless. Use of a random effects model does not make interpretation of analysis with extreme heterogeneity meaningful.

Thank you. We have adhered to the highest standard of conduct within this review, including decisions to perform meta-analysis, as informed by the Cochrane Handbook for Systematic Reviews of Interventions (Chapter 9).[2] Although statistical heterogeneity is a consequence of clinical and methodological diversity across studies, Cochrane does not recommend using the I^2 value to determine whether to conduct or present the findings of meta-analyses.[2] For this reason, we

considered clinical and methodological diversity across the studies for each outcome in deciding whether to pool. We chose only to pool the patient outcomes because it appeared that the participants, exposures, outcomes, study designs, and risk of bias were sufficiently similar across studies to make the analyses meaningful. Given substantial clinical and methodological heterogeneity, and the relatively low quality of many of the studies included for other outcomes, we believe that statistical pooling would not be reasonable nor add value.

Within the methods, we have added the thresholds for I^2 suggested in the Cochrane Handbook, as follows *"We considered an I^2 value of 0% to 40% to be low (potentially unimportant), 30% to 60% to be moderate, 50% to 90% to be substantial, and 75% to 100% to be considerable heterogeneity."* As suggested, we have updated the threshold for significance of the chi-square test to 0.10, although we have generally assumed that statistical heterogeneity will exist, thus relied primarily on the I^2 to provide an indication of the extent.

We agree that the random effects model does not make interpretation of an analysis with considerable heterogeneity more meaningful, and have included subgroup analyses in attempt to explore possible causes of heterogeneity for the analyses of intraoperative complications and length of stay. Since these were unable to explain the considerable heterogeneity present in the analyses, we have suppressed the point estimates within the results. We have instead provided a narrative summary, as follows: *"We found considerable between-study heterogeneity in the analyses for intraoperative complications (I^2 = 82%) and length of stay (I^2 = 86%), which could not be explained via subgroup analyses by procedure type, thus we have suppressed the average estimates of effect (findings of these analyses are shown in Supplementary file 5). For length of stay, the results of one study on cardiac surgeries favoured sleep deprived surgeons, while the others had null results. For intraoperative complications, the findings of one study favoured non-sleep deprived surgeons, but the others had null results."*

3. Please do not switch between the term 'heterogeneity' and 'homogeneity' to refer to the same phenomena. Failure to find statistical heterogeneity does not infer homogeneity. It just infers that you did not find statistical heterogeneity.

Thank you. We have removed the reference to homogeneity on page 7 of the revised manuscript.

4. This review remains excessively broad. A focus on either patient outcomes or clinician outcomes would improve this review, along with a pre-defined primary outcome. As it stands, I suggest all results should be discounted for multiple-comparisons.

We have previously commented (in response to the Associate Editor at revision #1) on the purpose for the broad scope of this review, which is primarily to synthesize the available evidence, raise awareness of the weaknesses within the current evidence base, and motivate higher quality research. We previously made substantial attempts within the manuscript (shown below, as presented in our previous response) to acknowledge the poor quality of the existing research and heterogeneity within the findings as a key result of the review, and to provide recommendations for future projects.

(a) The broad scope is justified within the introduction: “Given this void, we judged that a systematic review focusing broadly on primary research relevant to the Canadian context would be a fundamental starting point to examine the effects of fatigue and chronic sleep restriction on physicians in independent practice, and on interventions to combat these effects.”

(b) We have incorporated the key message prominently in the first paragraph of the discussion: “The key message gleaned from this review is that despite growing interest in the topic of physician wellness, the robust evidence needed to inform individual and systems-level fatigue management strategies is lacking.”

(c) We have modified the final paragraph of the discussion to reiterate our key message, which now reads as follows: “The most salient finding of this review is that the current evidence is insufficient to inform policy and practice. Accordingly, a 2016 research summit on physician wellness and burnout outlined the need for timely, relevant and methodologically robust research to inform practice and policy. The findings herein may be used as motivation for researchers and practitioners to develop and design methodologically strong research programs related to physician fatigue, inform successful research grant proposals, and lobby healthcare organizations to increase the focus on physician fatigue management programs. It will be important to make use of existing validated measures consistently in future research. Identifying outcomes of importance to physicians and their patients should be prioritized, such that these may be collected within intervention studies. Reporting these consistently will allow for the effective synthesis of findings and reduce research waste. Integrated knowledge translation strategies involving multiple stakeholder groups (e.g., physicians, patients, medical schools, physicians’ associations and governing bodies, policymakers) may help to ensure that the research is relevant and facilitates decision-making.”

References:

1. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003;327:557-560.

Higgins JPT, Green S (editors). The Cochrane handbook for systematic reviews of interventions, version 5.1.0. London, UK: The Cochrane Collaboration, 20

VERSION 3 – REVIEW

REVIEWER	Gordon S. Doig University of Sydney
REVIEW RETURNED	09-Jul-2018

GENERAL COMMENTS	Review of manuscript mjopen-2018-021967.R2 entitled "The impact of fatigue and insufficient sleep on physician and patient outcomes: A systematic review". 1. I understand that interpretation of p-value of heterogeneity alone has limited relevance. This is why I am requesting that you interpret it in addition to I2. Each time you report results of a meta-analysis in the abstract, manuscript or figures, please report p-value for heterogeneity immediately after you report I2. Please interpret both I2 and p-heterogeneity. 2. This review remains excessively broad. A focus on either patient outcomes or clinician outcomes would improve this review, along with a pre-defined primary outcome. As it stands, I suggest all results should be discounted for multiple-comparisons.

VERSION 3 – AUTHOR RESPONSE

Comments from Reviewer 3:

1. I understand that interpretation of p-value of heterogeneity alone has limited relevance. This is why I am requesting that you interpret it in addition to I2. Each time you report results of a meta-analysis in the abstract, manuscript or figures, please report p-value for heterogeneity immediately after you report I2. Please interpret both I2 and p-heterogeneity.

Thank you. We have added in the p-value for heterogeneity following our reports of the I^2 throughout the manuscript.

2. This review remains excessively broad. A focus on either patient outcomes or clinician outcomes would improve this review, along with a pre-defined primary outcome. As it stands, I suggest all results should be discounted for multiple-comparisons.

Based on comments from the Editor, we have not altered the scope of the review. However, we have addressed concerns regarding multiple comparisons by altering the "strengths and limitations of this study" to read (bullet point 2): *"The review was limited by the quality of the included studies, which was often poor. Confidence in our conclusions may be weakened due to multiple comparisons."* We have also added a short statement to the limitations (within the discussion on p.21): *"Confidence in the conclusions is limited due to multiple comparisons."* Finally, we have addressed the issue of multiple comparisons within the conclusion (p.21): *"Our overall confidence in the findings is low, owing to multiple comparisons and a body of research that is hindered by methodological weaknesses."*