PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Impact of intra-work rest breaks on doctors' performance and wellbeing: systematic review
AUTHORS	O'Neill, Aimee; Baldwin, David; Cortese, Samuele; Sinclair, Julia

VERSION 1 – REVIEW

REVIEWER	Allan, Julia University of Aberdeen, Aberdeen Health Psychology Group
REVIEW RETURNED	18-Jun-2022

GENERAL COMMENTS	This is a timely review of an interesting, policy relevant area. As the authors clearly identify, workload pressures on doctors are high and increasing, yet the healthcare sector is generally recognised to be behind other high risk industries (e.g. aviation, offshore drilling, military etc) in its approach to fatigue risk management. An up to date summary of the evidence relating to the impact of breaks on doctors' performance and wellbeing will be of interest to many. The introduction clearly and concisely justifies the review. The methods are appropriate and well reported. The results, while extremely heterogenous and therefore difficult to synthesise, provide a useful overview of an important area. As such, I would support publication of this manuscript and have only minor suggestions. 1. Additional clarity around what were considered to be 'wellbeing' and 'performance' outcomes, and how any definitions / classification criteria arose would be welcome. Both are extremely broad terms - wellbeing could encapsulate everything from momentary fatigue to general quality of life, job satisfaction etc, and performance could be conceptualised as anything from meeting of performance targets, processing speed, accuracy, guideline compliance, safety violations, patient outcomes etc. 2. It would be interesting to reflect in the discussion on the likely impact of the content of breaks on the outcomes of interest – whether doctors were sleeping, eating, spending time in activities vs resting, etc. 3. The search strategy appears comprehensive. Related to point 1 though, how were the performance and wellbeing terms generated to ensure that the search picked up papers relevant to such diverse, multifaceted concepts? Also, please include the range of publication years included in the search. 4. Please report the agreement between first and second screeners, and expand on the second author's role in extraction (did they independently extract a % and compare, or did they verify all

	extraction?) 5. P7, line 43/44 – 'hiding ignorance' would benefit from further explanation at this point as it's not immediately clear what this is in the context of a study outcome .It is explained later in the paper, but would be useful to readers to clarify on first mention. 6. Eligibility criteria state that only studies relating to qualified medical doctors were included but the results describe at least one study about trainees (e.g [46]). Clarify for consistency. 7. In discussion of the risk of bias, it might be worth explicitly acknowledging the low feasibility of blinding in this context.

REVIEWER	Valentin, Jan Aalborg University Hospital, Department of Clinical Medicine
REVIEW RETURNED	21-Jun-2022

GENERAL COMMENTS	Thank you for allowing me to review the manuscript "Impact of intra-work rest breaks on doctors' performance and wellbeing: systematic review". The study touches upon an important topic within the quality of health care and I agree with the conclusion reached by the authors. The systematic search seems thoroughly conducted and the search strings are are well presented and easy to follow. However, the manuscript lacks detailed overview of results and it is difficult to see how the authors reached their conclusion without reading the studies included in the review. Major concerns: After reviewing a subset of the original research papers included in the systematic review review, I agree with authors that the results are highly inconsistent. However, I would appreciate if the results of the quantitative studies were to be presented much more systematic in figures with confidence intervals and standardized effects if possible. This would give a better overview of the inconsistency and would allow for assessment of whether lack of evidence was related to lack of power or lack of relevance. Furthermore, results should also be stratified by outcome type (wellbeing and performance). Table 1-5 could easily be put in supplementary to make room for other tables and figures. The JBI checklist for assessing Risk of bias in Observational cohort study doesn't seem appropriate for studies with continues exposure. The authors have been quick to Q1 and Q2 for N/A in all studies of table 3. However, these items also apply for continuous exposures (or multiple exposure levels). E.g. groups defined by the levels of the exposure (continuous or not) should be similar and recruited from the same population. Moreover, the JBI checklists are much less detailed than ROBINS, thus additional comments on the risk of bias, when the JBI checklists are used is appreciated. Looking at Supplementary table 1, it is clearly difficult to give a systematic description of the outcomes used in the included studies, however, I would like to know what the authors believe constitute measures of wellbeing measures and what constitute measures of performance. Moreover, I find it concerning that not a single study used a generic measure of well-being, such as WHO5 or EQ-5d. Similarly, I find it concerning that performance was not measured by quality of health process indicators in any of the included studies. Were these studies excluded? If not, then the discussion should

	entail this concern, otherwise such studies should be included. Moreover, it should be possible to do some kind of grouping of the outcomes measures and present tabulations of these groups. Other concerns: Page 4, line 7: This is neither a strength nor a limitation. Page 4, line 9: I am not sure this is true, since I did not see a multilanguage search string. Page 5, line 15: How is this relevant for the study? Page 5, line 46: I would like to know under which circumstances an investigation of breaks on well-being/performance does not necessitate a comparator? Studies using qualitative methods? Page 12, line 40: How is this performance? As mentioned above, what is meant by performance and wellbeing should be clearly defined. Page 20, line 39: This sentence should be rephrased, since I believe such investigations are possible without reforming and standardising definitions of breaks and performance. Thus, it should not be phrased as a necessity.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Dr. Julia Allan, University of Aberdeen

Comments to the Author:

This is a timely review of an interesting, policy relevant area. As the authors clearly identify, workload pressures on doctors are high and increasing, yet the healthcare sector is generally recognised to be behind other high risk industries (e.g. aviation, offshore drilling, military etc) in its approach to fatigue risk management. An up to date summary of the evidence relating to the impact of breaks on doctors' performance and wellbeing will be of interest to many. The introduction clearly and concisely justifies the review. The methods are appropriate and well reported. The results, while extremely heterogenous and therefore difficult to synthesise, provide a useful overview of an important area. As such, I would support publication of this manuscript and have only minor suggestions.

Thank you for these positive and encouraging comments.

1. Additional clarity around what were considered to be 'wellbeing' and 'performance' outcomes, and how any definitions / classification criteria arose would be welcome. Both are extremely broad terms - wellbeing could encapsulate everything from momentary fatigue to general quality of life, job satisfaction etc, and performance could be conceptualised as anything from meeting of performance targets, processing speed, accuracy, guideline compliance, safety violations, patient outcomes etc.

Our search was deliberately broad and does indeed include all of these outcomes. We mention in the text that we used Health Sciences Descriptors (DeCS) and Medical Subject Headings (MeSH) and we have also added the following paragraph to the main text:

> "Wellbeing" and "work performance" are broad constructs that lack a single definition. As we aimed to be as broad as possible in this search, capturing the breadth of research in the field, we wished to encompass as many working definitions of these constructs as possible through a comprehensive list of search terms. Wellbeing outcomes referred to any measures of, or related to, mental health, physical health and quality of life. Work performance included any measures of, or related to, clinicians' ability to carry out their duties, such as errors, adverse

events, appraisals, patient feedback, quality of care, revalidation, ability to meet targets, and so forth. Outcomes relating to wellbeing and work performance also often overlap (e.g. sickness absence, perceived stress) therefore it was not our intention to divide the two constructs but rather to be inclusive of any papers investigating either, or both, outcomes. We referred to research papers in the fields of occupational wellbeing and job performance to gather the extensive list of terms and a subject librarian was consulted throughout to ensure the comprehensiveness of the search."

2. It would be interesting to reflect in the discussion on the likely impact of the content of breaks on the outcomes of interest – whether doctors were sleeping, eating, spending time in activities vs resting, etc.

We have added the following paragraph to the Discussion:

> "It is unclear whether the content or duration of breaks is more important than the ability to take one. There were mixed findings for standard 30-minute uninterrupted breaks, though this is likely due to a lack of studies on the topic. Naps, microbreaks, and yoga and mindfulness interventions showed improvement to wellbeing and/or performance outcomes. However, outcome measures were dissimilar across studies, preventing valid comparison of break interventions and durations."

3. The search strategy appears comprehensive. Related to point 1 though, how were the performance and wellbeing terms generated to ensure that the search picked up papers relevant to such diverse, multifaceted concepts?

See our response to point #1.

Also, please include the range of publication years included in the search.

The search was not limited by date, however, records meeting the inclusion criteria were from 2002 and later. We have added the following text to the Results section: "The records meeting the inclusion criteria ranged in publication date from the year 2002 to 2021"

4. Please report the agreement between first and second screeners,

We have added the following bold text in the Study Selection section: "Each study title and/ or abstract was assessed by two independent assessors against the inclusion criteria, **with an agreement rate of 98.2%**."

…and expand on the second author's role in extraction (did they independently extract a % and compare, or did they verify all extraction?)

We have added the following to the 'Data extraction and analysis' section: "Data extraction for each study was completed by the primary author using a standardised table and **all data extraction was** verified by a second, senior author **throughout the extraction process**."

5. P7, line 43/44 – 'hiding ignorance' would benefit from further explanation at this point as it's not immediately clear what this is in the context of a study outcome .It is explained later in the paper, but would be useful to readers to clarify on first mention.

We have moved the definition text to the first instance of these terms.

6. Eligibility criteria state that only studies relating to qualified medical doctors were included but the results describe at least one study about trainees (e.g [46]). Clarify for consistency.

Despite being fully qualified (i.e. no longer medical students), junior doctors are often referred to as trainees until they are consultants. Indeed the term 'trainee' means they are still doing some level of training (could be foundation level, registrars, core trainees), though working as fully qualified doctors from a legal standpoint.

We have added the following to the main text: “Some papers might refer to junior doctors as ‘trainees’. Despite being fully qualified, this is a common term for doctors who are not yet consultants.”

7. In discussion of the risk of bias, it might be worth explicitly acknowledging the low feasibility of blinding in this context.
We have added the following text to the Discussion:

> “While the feasibility of blinded experiments in break-taking research is low and unlikely, there is scope to reduce confounding and introduce more randomised control trials in this area.”

Reviewer: 2
Dr. Jan Valentin, Aalborg University Hospital

Comments to the Author:
Thank you for allowing me to review the manuscript "Impact of intra-work rest breaks on doctors’ performance and wellbeing: systematic review". The study touches upon an important topic within the quality of health care and I agree with the conclusion reached by the authors. The systematic search seems thoroughly conducted and the search strings are are well presented and easy to follow. However, the manuscript lacks detailed overview of results and it is difficult to see how the authors reached their conclusion without reading the studies included in the review.

Major concerns:

1) After reviewing a subset of the original research papers included in the systematic review review, I agree with authors that the results are highly inconsistent. However, I would appreciate if the results of the quantitative studies were to be presented much more systematic in figures with confidence intervals and standardized effects if possible.
This would give a better overview of the inconsistency and would allow for assessment of whether lack of evidence was related to lack of power or lack of relevance.

We considered this issue very carefully when we planned the protocol of the systematic review and after conducting the scoping search. Each study used very different outcomes and statistical approaches: they also have multiple measures of wellbeing and/or performance, with very little overlap between studies. Therefore, presenting (available) effect sizes with relative confidence intervals in the same figure could be misleading for the reader. Indeed, quantitative meta-analysis was not possible because of the marked heterogeneity of the data. Furthermore, statistical details and confidence intervals were not included in the results Table due to its existing length and complexity.

2) Furthermore, results should also be stratified by outcome type (wellbeing and performance). Table 1-5 could easily be put in supplementary to make room for other tables and figures.

Because each study can have both wellbeing and performance outcomes (inclusion criteria states “and/or” for the two terms), to do this we would need to split results both between and within studies. Whilst not impossible, this would reduce, rather than improve, clarity. Additionally, as wellbeing and performance may be quite intricately linked, it can be difficult to separate the two at times. For example, sickness absence, perceived stress or burnout could arguably be classified under both types of outcome, as they are directly related to duties carried out on the job but are also wellbeing issues. It was not our intention to split the review into two parts. Furthermore, 14 out of 32 studies contained both wellbeing and performance outcome measures, some of which are unable to be separated as the two outcomes are either correlated with one another, or in qualitative studies sometimes occurred within a single sentence. It is our hope that the new paragraph on outcome definition will help clarify this for readers.

3) The JBI checklist for assessing Risk of bias in Observational cohort study doesn't seem appropriate for studies with continues exposure. The authors have been quick to Q1 and Q2 for N/A in all studies of table 3. However, these items also apply for continuous exposures (or

multiple exposure levels). E.g. groups defined by the levels of the exposure (continuous or not) should be similar and recruited from the same population.

As the Reviewer might be aware that, while for randomized controlled trials (RCTs) the Cochrane risk of bias tool 2 (RoB2) is the most commonly used and accepted tool and, for non-randomised trials of interventions, ROBINS-I is becoming popular, there is no established/agreed tool to assess observational studies/non-intervention studies. We considered the Newcastle-Ottawa Scale not suitable, given the design of a number of studies included in the review (e.g. one-group post-test only design using qualitative survey evaluation or single occasion focus groups). We acknowledge that the JBI is not ideal and that this and other tools will continue to evolve, but we are also aware that it is common practice to adapt the tools (e.g. dropping some items which are not relevant) and that the key requirement is to present the choices and the ratings/judgement in a transparent way, as we have done. We have carefully reviewed our judgments and, in relation to point #4, clarified the rationale underpinning our judgment for each item in the JBI checklists, when applicable.

4) Moreover, the JBI checklists are much less detailed than ROBINS, thus additional comments on the risk of bias, when the JBI checklists are used is appreciated.

We were unsure of the type of additional comments requested here. We have however provided our rationale for scoring 'Yes', 'No', 'Unclear', or 'N/A' on the JBI checklists in Supplementary Tables 2-4. This is also referenced in the main text: "As JBI checklists contain less detail than Cochrane tools, our rationale for JBI ratings are given in Supplementary Tables 2-4."

If the additional comments requested here instead refer to quality assessments, we have corrected the wording in this section as PRISMA highlights the need for risk of bias tools rather than quality assessments. We had indeed undertaken risk of bias assessments, instead of quality assessments. Otherwise, if the reviewer is instead requesting a composite score (overall risk of bias assessment e.g. low, moderate, high), this is not advised for this type of scale/checklist, as they were not designed for this purpose. While several studies have done this, it is not advised by the Cochrane Collaboration, which advocates for scores on specific domains or components, rather than composite scores which are difficult to interpret.

5) Looking at Supplementary table 1, it is clearly difficult to give a systematic description of the outcomes used in the included studies, however, I would like to know what the authors believe constitute measures of wellbeing measures and what constitute measures of performance.

We have added the following paragraph to the main text:

> "Wellbeing" and "work performance" are broad constructs that lack a single definition. As we aimed to be as broad as possible in this search, capturing the breadth of research in the field, we wished to encompass as many working definitions of these constructs as possible through a comprehensive list of search terms. Wellbeing outcomes referred to any measures of, or related to, mental health, physical health and quality of life. Work performance included any measures of, or related to, clinicians' ability to carry out their duties, such as errors, adverse events, appraisals, patient feedback, quality of care, revalidation, ability to meet targets, and so forth. Outcomes relating to wellbeing and work performance also often overlap (e.g. sickness absence, perceived stress) therefore it was not our intention to divide the two constructs but rather to be inclusive of any papers investigating either, or both, outcomes. We referred to research papers in the fields of occupational wellbeing and job performance to gather the extensive list of terms and a subject librarian was consulted throughout to ensure the comprehensiveness of the search."

6) Moreover, I find it concerning that not a single study used a generic measure of well-being, such as WHO5 or EQ-5d. Similarly, I find it concerning that performance was not measured by quality of health process indicators in any of the included studies. Were these studies excluded? If not, then the discussion should entail this concern, otherwise such

studies should be included.

We did not exclude this type of study as it was the purpose of the search to retrieve these studies. There are simply very few studies whichincluded such a measure. A few (e.g. Berastegui 2020, Nitzsche 2017, Kalboussi 2020, Babbar 2019) used instruments such as the Maslach Burnout Inventory (MBI), but overall use of validated instruments was low. In response to the raised concern, we have added the following to the Discussion: "Indeed the use of validated instruments for measures of wellbeing or performance was low overall."

7) Moreover, it should be possible to do some kind of grouping of the outcomes measures and present tabulations of these groups.

Please see our response to point #2.

Other concerns:

8) Page 4, line 7: This is neither a strength nor a limitation.

Noted. This has been removed from the manuscript.

9) Page 4, line 9: I am not sure this is true, since I did not see a multilanguage search string.

To clarify, we state: "No limits were placed on design, country or language to ensure a comprehensive review of the subject area" Although the search string was in English, the resultwere returned in multiple languages. Language was not a reason for exclusion. E.g. Kalboussi et al. 2020, an included paper, is in French. Excluded studies included papers in German, Spanish, and Czech, among others.

10) Page 5, line 15: How is this relevant for the study?

We were required to insert this paragraph here as part of the requirements for submission to this journal.

Page 5, line 46: I would like to know under which circumstances an investigation of breaks on well-being/performance does not necessitate a comparator? Studies using qualitative methods?

We did not require an intervention or two groups for a study to be included in our review, as we kept our search as broad as possible. When stating "No comparator" this could indeed include qualitative studies as well as observational studies with only 1 group and no intervention. While most quantitative studies in this field have some type of intervention, this was not required for inclusion in our review.

Page 12, line 40: How is this performance? As mentioned above, what is meant by performance and wellbeing should be clearly defined.

It is our hope that by providing more information on the definitions, as requested above, the reasons for inclusion might be more apparent. The screeners felt that *"'sharing ignorance' (detecting and sharing unknown knowledge and learning from failures) and 'hiding ignorance' (deliberately preventing knowledge sharing)"* affected the execution of doctors' duties and therefore related to performance of the job/job performance (per the definition). The consensus decision was that as much of medicine is 'apprentice based', this kind of learning (or lack thereof) is likely to have an impact on performance.

Page 20, line 39: This sentence should be rephrased, since I believe such investigations are possible without reforming and standardising definitions of breaks and performance. Thus, it should not be phrased as a necessity.

Our original sentence was as follows: “To properly understand the effectiveness of breaks for doctors and justify financial and organisational investment in break facilitation, a panoply of policy makers, regulators and research bodies need to agree the priorities so that the evidence base can be developed quickly and effectively.”

We have amended the wording to emphasise that it is not a necessity. However, whilst it is not necessary that we set priorities for the research to be *possible*, it is our view that this is not helping to further the field and it is certainly taking longer than if there was some consensus. The results of this systematic review show that there is no consensus on any outcome measurement, timing of breaks, definition of what constitutes a break, and so forth. This allows no valid comparison or meta-analysis and means we are all ‘singing from different hymn sheets’ when it comes to determining whether breaks are beneficial or otherwise to doctors’ wellbeing and/or performance of their jobs. As this was the purpose of the review, we concluded that there is inconsistency and that it would be beneficial to have some priority setting around this to encourage greater consistency in investigations.

VERSION 2 – REVIEW

REVIEWER	Allan, Julia University of Aberdeen, Aberdeen Health Psychology Group
REVIEW RETURNED	25-Oct-2022

GENERAL COMMENTS	The authors have fully addressed all of the points I raised in my initial review.

REVIEWER	Valentin, Jan Aalborg University Hospital, Department of Clinical Medicine
REVIEW RETURNED	24-Oct-2022

GENERAL COMMENTS	I acknowledge the difficulties in synthesizing the results, and have no further comments. The authors may consider adding "and evidence mapping" to the end of the title.