PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	The impact of leadership behavior on physician wellbeing, burnout, professional fulfillment and intent to leave: a multi-center cross-sectional survey study
AUTHORS	Mete, Mihriye; Goldman, Charlotte; Shanafelt, Tait; Marchalik, Daniel

VERSION 1 – REVIEW

REVIEWER	Messias, Erick UAMS Medical Center
REVIEW RETURNED	15-Nov-2021

GENERAL COMMENTS	This paper examines the role of leadership on three main outcomes: professional fulfillment, burnout, and intention to leave. Leadership was measured using the revised 9-item Mayo Clinic Participatory Management Leadership Index; professional fulfillment and burnout was measured using the Professional Fulfillment Index – 6 items for fulfillment, 10 items for burnout (combining 4 items for the assessment of work exhaustion and 6 items to assess interpersonal disengagement). Intent to leave was ascertained using one item: What is the likelihood that you will leave your institution within two years? The authors also assessed sleep using the PROMIS short-form sleep-related impairment scale – 8 items. The authors divided the leadership measurements into terciles and found that as the leadership assessment improved professional fulfillment increased while professional burnout and intention to leave decrease. These findings highlight the impact leadership can have on professional wellbeing and organization stability. Suggestions and recommendations: 1. It is not clear how much the sleep measures add to the findings – if anything could be distracting. I would suggest either removing it altogether or incorporating it fully to the discussion – favoring removing it altogether. It could be a separate paper or report on leadership impact on sleep? 2. Table 2 displays the terciles as 1, 2, and 3. Would suggest naming these so it is clear when reading it which tercile is high, average, low leadership behavior. Same follows for the figures. 3. Recommend adding the items in the leadership index as part of the appendix. It is available in the Proceedings paper. 4. Is there a way to report the amount of correlation – albeit inverse correlation – between fulfillment and burnout? Same for burnout and intention to leave. It would be logical to think there is significant collinearity between these constructs. 5. The three figures convey limited information and the labelling is somewhat confusing. For example: on Table 2 the percentage of

	those having professional fulfillment on the third tercile is 47% while on figure 1, the “likelihood” of professional fulfillment is 57% for women and 71% for men. Assistant is needed in reconciling these estimates. Similar discrepancies appear in the burnout prevalence and in the percent intending to leave. This is likely the result of modelling so some explanation and labelling – in particular of the figures – will help. For example, on the label for figure 1 reading “Likelihood (%, 95% CI) of Professional Fulfillment Status by the Terciles of Unit-Level” – you may add “adjusting for….” 6. Still on modelling, age seems to be a significant factor in these outcomes so it would be informative to have information on age and its effects on these measures – assuming the authors have information on age of respondents. 7. Regarding limitations the authors report the response rate of 45% - as potential response and recall bias. What would be more impactful in this situation would be the problem of selection bias of responders – i.e. those burned out are more likely to respond to surveys about burnout. 8. Another potential limitation is the burnout measure used which seems to contemplate only two of the three proposed components of burnout. 9. Finally, it might be helpful to report on these findings separating academic and non-academic healthcare setting – or eventually report on academic settings separately. Or run a sensitivity analysis across the 11 sites reporting so see if there are specific sites bringing up specific issues – such as gender bias effects. It is possible that confidentiality issues prevent this recommendation. 10. There seems to be significant differences across specialties so the author may touch on it or work on ways to categorize these specialties – say into groups of at least 100 respondents – so that the differences across these specialties groups can be highlighted and understood. I hope these comments are helpful and I commend the authors for working on such important problem affecting our healthcare teams.

REVIEWER	Mohr, David VA Boston Health Care System Jamaica Plain Campus
REVIEW RETURNED	16-Nov-2021

GENERAL COMMENTS	In the introduction, it would help to specify what type of leadership behaviors are being measured as part of the research background and design. Are these “transactional” behaviors, empowering behaviors, servant leadership, laissez-faire, abusive behaviors, or something else? It would be important to define this in terms of knowing how to focus training to develop these skillsets for leaders. It becomes a bit clearer in the methods what the items focus on but the focus of the paper would benefit by trying to describe or label this particular type of leadership style. The inclusion of sleep impairment was an interesting choice to highlight in the abstract. and in the study results. I would like to see a bit more about this variable and research on it in relation to burnout. It likely is important but some conceptual background in the introduction would help, and could also highlight the contribution to the literature by including this.

	Has the intent to leave item been studied/validated before? The two-year time frame seems a bit longer than what is usually used on some surveys. Table 2 – it would help to provide the cutpoints used for the leadership behavior score tertiles. Author's may want to present a Cramer's V statistic or similar to provide a measure of the strength of association as everything will likely be significant with the large sample size. Authors provide a similar measure in the results when discussing the correlation coefficients. Authors note different findings with gender and its impact in the workplace and leadership roles. Authors may want to consider asking about the gender of the leader who is being rated in future surveys to help address some of the issues (indirectly) they note in the Discussion section. I was thinking there would be some mention for the "sleep" findings in the Discussion. It seemed like it was included in the survey for some other set of reasons but not given very much attention in the text of the manuscript. The work could benefit by either removing it to avoid being distracting or to embrace it by adding more explanation around the choice for inclusion and interpretation of findings. A strength of the study was the large sample size and institutions involved. A fixed effects model may be helpful to consider – if authors can match the respondent to institution. There are likely some systematic differences in burnout and fulfillment by study site.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

Thank you very much for taking the time to review our manuscript. We appreciate your suggestions and recommendations. Please see our responses below to your well-taken points.

1. It is not clear how much the sleep measures add to the findings – if anything could be distracting. I would suggest either removing it altogether or incorporating it fully to the discussion – favoring removing it altogether. It could be a separate paper or report on leadership impact on sleep?

We included sleep-related impairment in our models because of its strong association with our outcome measures, especially with burnout (r=0.55, p<0.001). Its correlation with professional fulfillment is -0.33 (p<0.001). Overall, its inclusion increases the predictive power of the logistic regression models for burnout and professional fulfillment. Please note the decrease in AUC statistics in Table 3. However, it has a weak correlation with leadership behavior (r=-0.14, p<0.001) and, therefore, is, statistically, not a confounder in the model. Our updated analyses confirmed that the effect of leadership on the outcomes was not affected by the exclusion of the sleep-related impairment variable. The revised table present the results of the models without sleep-related impairment variable. Our results section was also revised accordingly. Thank you for this suggestion.

1. Table 2 displays the terciles as 1, 2, and 3. Would suggest naming these so it is clear when reading it which tercile is high, average, low leadership behavior. Same follows for the figures.

This clarification has been made.

1. Recommend adding the items in the leadership index as part of the appendix. It is available in the Proceedings paper.

The leadership index was added as an appendix with an acknowledgment of Mayo Copyright and permissions.

1. Is there a way to report the amount of correlation – albeit inverse correlation – between fulfillment and burnout? Same for burnout and intention to leave. It would be logical to think there is significant collinearity between these constructs.

Yes, burnout and professional fulfillment are strongly correlated (r=-0.65, p<0.001) and there is a reciprocal relationship between them (causality not clear). That is the reason, we did not use one to estimate the other (Models 1 and 2). Burnout is also strongly related to intent to leave (Model 3: OR=2.4, 95% CI:2.2-2.7, p<0.001) and so is professional fulfillment (Model 3: OR=0.45, 95% CI:0.40-0.52, p<0.001). Despite the high correlation between professional fulfillment and burnout, the model 3 doesn't suffer from multicollinearity problem based on our model checks. Both burnout and professional fulfillment are independently associated with the intent to leave outcome.

1. The three figures convey limited information and the labelling is somewhat confusing. For example: on Table 2 the percentage of those having professional fulfillment on the third tercile is 47% while on figure 1, the "likelihood" of professional fulfillment is 57% for women and 71% for men. Assistant is needed in reconciling these estimates. Similar discrepancies appear in the burnout prevalence and in the percent intending to leave. This is likely the result of modelling so some explanation and labelling – in particular of the figures – will help. For example, on the label for figure 1 reading "Likelihood (%, 95% CI) of Professional Fulfillment Status by the Terciles of Unit-Level" – you may add "adjusting for…."

We apologize for this confusion. Table 2 percentages reflect row percentages to make the interpretation easier for the descriptive statistics. For professional fulfillment, percentages in Table 2 are telling us what percent of those with high professional fulfillment are in each tertile. If we presented column percentages, it would have been more difficult to assess the differences between specialties. For instance, if a given specialty is similar to overall distribution, the row percentages would be roughly distributed around 33% for each tertile. However, if they are 14%, 30%, 56% like in dermatology, we can infer that the leadership evaluations for dermatology tend to be higher on average. If we were to present column percentages, we would see the percentage from each specialty within each tertile.

This may be a matter of preference. We are willing to switch to column percentages if requested. The figures represent the likelihood of having professional fulfillment for each tertile, corresponding to column percentages, since, in the models, we are examining how the leadership tertiles are related to the probability of having professional fulfillment, being burned out or intending to leave. Yes, they are based on adjusted models but that is not the reason why they were different from those percentages in Table 2. We hope that this clarifies the interpretation of the results and the discrepancies between figures and tables.

1. Still on modelling, age seems to be a significant factor in these outcomes so it would be informative to have information on age and its effects on these measures – assuming the authors have information on age of respondents.

There is evidence that prevalence of burnout is smaller among older physicians. For instance, Table3 below is taken from Shanafelt et al. 2019 and shows that physicians in the 65+ age group are less likely to have burnout compared to those in <35 y. (OR=0.44). Similarly, satisfaction with work life integration (WLI) increases with age.
As you rightly point out, the age of physicians is a relevant variable to be controlled for in models. However, in the first dataset distributed by PWAC (PWAC 1.0), age was not provided to the investigators to avoid potential loss of confidentiality especially for smaller specialties. Age along with gender and specialty can easily help reveal the identity of physicians in small specialties like urology, ophthalmology etc.
Age would likely to be a confounder if it is simultaneously related to leadership evaluation and to a given outcome. Unfortunately, we can’t evaluate that in our dataset. We don’t know if physicians become less critical of leadership as they get older, everything else being constant. We admit that it is a limitation and added a few lines to the discussion regarding this concern.

Shanafelt TD, West CP, Sinsky C, Trockel M, Tutty M, Satele DV, Carlasare LE, Dyrbye LN. Changes in Burnout and Satisfaction With Work-Life Integration in Physicians and the General US Working Population Between 2011 and 2017. Mayo Clin Proc. 2019 Sep;94(9):1681-1694. doi: 10.1016/j.mayocp.2018.10.023. Epub 2019 Feb 22. PMID: 30803733.

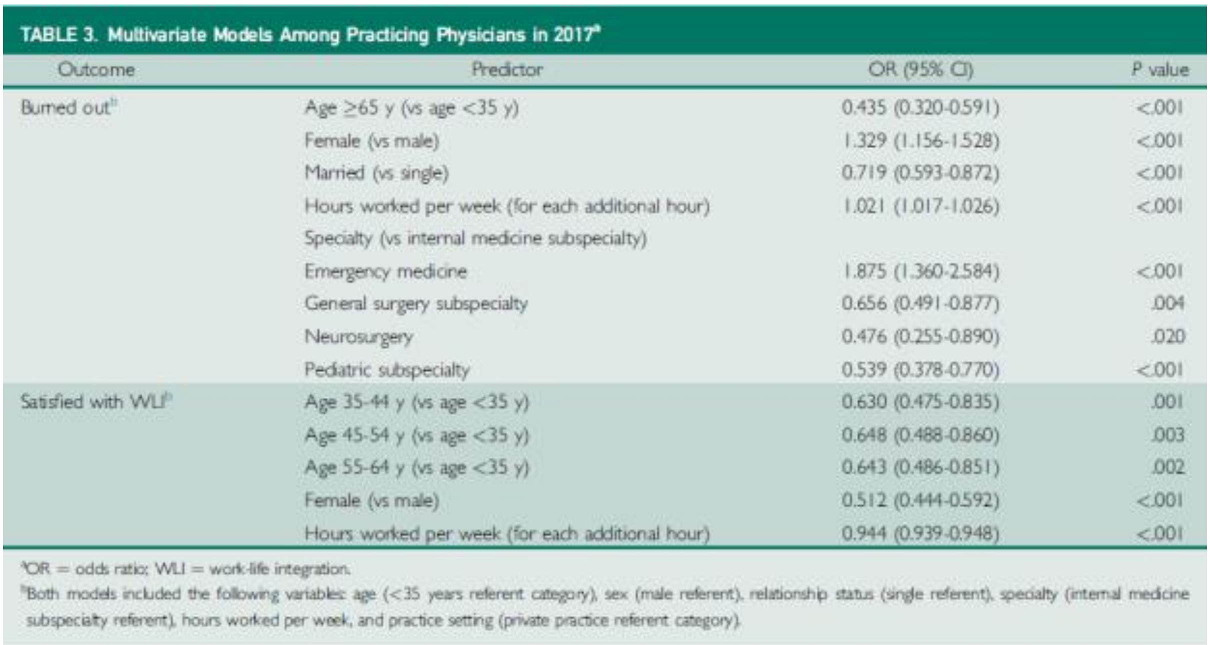

TABLE 3. Multivariate Models Among Practicing Physicians in 2017[a]

Outcome	Predictor	OR (95% CI)	P value
Burned out[b]	Age ≥65 y (vs age <35 y)	0.435 (0.320-0.591)	<.001
	Female (vs male)	1.329 (1.156-1.528)	<.001
	Married (vs single)	0.719 (0.593-0.872)	<.001
	Hours worked per week (for each additional hour)	1.021 (1.017-1.026)	<.001
	Specialty (vs internal medicine subspecialty)		
	Emergency medicine	1.875 (1.360-2.584)	<.001
	General surgery subspecialty	0.656 (0.491-0.877)	.004
	Neurosurgery	0.476 (0.255-0.890)	.020
	Pediatric subspecialty	0.539 (0.378-0.770)	<.001
Satisfied with WLI[b]	Age 35-44 y (vs age <35 y)	0.630 (0.475-0.835)	.001
	Age 45-54 y (vs age <35 y)	0.648 (0.488-0.860)	.003
	Age 55-64 y (vs age <35 y)	0.643 (0.486-0.851)	.002
	Female (vs male)	0.512 (0.444-0.592)	<.001
	Hours worked per week (for each additional hour)	0.944 (0.939-0.948)	<.001

[a] OR = odds ratio; WLI = work-life integration.
[b] Both models included the following variables: age (<35 years referent category), sex (male referent), relationship status (single referent), specialty (internal medicine subspecialty referent), hours worked per week, and practice setting (private practice referent category).

1. Regarding limitations the authors report the response rate of 45% - as potential response and recall bias. What would be more impactful in this situation would be the problem of selection bias of responders – i.e. those burned out are more likely to respond to surveys about burnout.

Low response rate, indeed, increases the concern of response bias and. it is an inherent problem of survey research targeting extremely busy professionals. Although 45% is not ideal, it is relatively high for physician surveys. It is unknown whether physicians who are burned out are more likely to participate in surveys (out of interest in the topic) or less likely to do so (due to apathy/being overwhelmed). A number of published studies using robust secondary surveys of non-responders and other approaches to assess response bias have, however, found that there is not a such a bias in either direction. That established evidence and our relatively high response rate for a physician survey gives us confidence that responders in our sample is representative of overall physician population unless all surveys overestimate burnout. Nonetheless, we overtly acknowledge this important limitation in the manuscript and have changed the relevant sentence in the discussion in keeping with the reviewer's suggestion.

1. Another potential limitation is the burnout measure used which seems to contemplate only two of the three proposed components of burnout.

We disagree with this point. The Stanford professional fulfillment index (PFI) has emotional exhaustion (4 items), interpersonal disengagement (6 items), and professional fulfillment (6 items) which are consistent with other burnout assessments. This instrument is a well-validated and widely used assessment. PFI burnout measures correlated highly with their closest related MBI equivalents (https://pubmed.ncbi.nlm.nih.gov/29196982/).   The emotional exhaustion and interpersonal disengagement domains have also been cross walked against other commonly used burnout instruments (such as the Maslach Burnout Inventory) in large samples of physicians:

- Brady et al. Establishing Crosswalks Between Common Measures of Burnout in US Physicians. J Gen Intern Med. 2021 Mar 31. Online ahead of print. https://pubmed.ncbi.nlm.nih.gov/33791938/

1. Finally, it might be helpful to report on these findings separating academic and non-academic healthcare setting – or eventually report on academic settings separately. Or run a sensitivity analysis across the 11 sites reporting so see if there are specific sites bringing up specific issues – such as gender bias effects. It is possible that confidentiality issues prevent this recommendation.

This is another great point. But we do not have an identifier for the institutions in our dataset due to confidentiality issues. all member organizations are either traditional University-based Academic Medical Centers or institutions affiliated with a University and/or which train residents and perform research.

1. There seems to be significant differences across specialties so the author may touch on it or work on ways to categorize these specialties – say into groups of at least 100 respondents – so that the differences across these specialty groups can be highlighted and understood.

We agree that it is somewhat taxing to compare the specialties in terms of their leadership evaluations. We thought that comparing the row percentages across specialties for a given tertile would be informative as in Table 2. A pairwise comparison of specialties, though, would require too many comparisons and create methodologic concerns. We have already grouped some sub-specialties together. For instance, surgery or medicine represent multiple sub-specialties. The smallest specialties (less than 100 participants) under the current grouping are Dermatology and Radiation Oncology, which were distinct enough to be their own categories. Furthermore, our goal in this study was not to compare specialties in terms of their leadership evaluations but to make sure the relationship between the leadership behavior score and burnout, PFI and intent to leave persists after adjusting for specialty. But we agree that it is informative to see the data and that is why we presented it in our paper.

I hope these comments are helpful and I commend the authors for working on such important problem affecting our healthcare teams.

We thank you very much for your constructive comments and helpful suggestions, which have substantially improved the manuscript.

Reviewer: 2

1. In the introduction, it would help to specify what type of leadership behaviors are being measured as part of the research background and design. Are these “transactional” behaviors, empowering behaviors, servant leadership, laissez-faire, abusive behaviors, or something else? It would be important to define this in terms of knowing how to focus training to develop these skillsets for leaders. It becomes a bit clearer in the methods what the items focus on but the focus of the paper would benefit by trying to describe or label this particular type of leadership style.

The Mayo Clinic Participatory Management Leadership Index evaluates leadership behaviors that engage and empower individuals to perform to the best of their ability. It also evaluates behaviors related to the concept of Wellness Centered Leadership, which includes dimensions of interpersonal power in leadership, situational leadership, transformational leadership. (https://pubmed.ncbi.nlm.nih.gov/33394666/) This instrument is a well-established tool that has been used across multiple centers and studies (examples below):

- https://pubmed.ncbi.nlm.nih.gov/25796117/
- https://pubmed.ncbi.nlm.nih.gov/32247343/
- https://pubmed.ncbi.nlm.nih.gov/34538425/
- https://pubmed.ncbi.nlm.nih.gov/33560424/
- https://pubmed.ncbi.nlm.nih.gov/33560424/
- https://pubmed.ncbi.nlm.nih.gov/32520754/

1. The inclusion of sleep impairment was an interesting choice to highlight in the abstract. and in the study results. I would like to see a bit more about this variable and research on it in relation to burnout. It likely is important but some conceptual background in the introduction would help, and could also highlight the contribution to the literature by including this.

Please see our response to Reviewer 1’s comment on the inclusion of the sleep impairment. The relationship between sleep impairment and burnout has been established (https://pubmed.ncbi.nlm.nih.gov/33284339/). We decided to remove this variable from the models as its inclusion does not change our conclusions about the effect of leadership evaluation on the outcomes.

1. Has the intent to leave item been studied/validated before? The two-year time frame seems a bit longer than what is usually used on some surveys.

The intent to leave item used in this study has been used in extensive prior surveys, for example:
https://pubmed.ncbi.nlm.nih.gov/29101932/
https://pubmed.ncbi.nlm.nih.gov/21356491/

Responses to this item have also been shown to correlate with future actual/objective departures:
https://pubmed.ncbi.nlm.nih.gov/30477483/

1. Table 2 – it would help to provide the cutpoints used for the leadership behavior score tertiles. Author's may want to present a Cramer's V statistic or similar to provide a measure of the strength of association as everything will likely be significant with the large sample size. Authors provide a similar measure in the results when discussing the correlation coefficients.

Thank you for this suggestion. We should have done this in the earlier version. We have now added the mean (SD) leadership score to Table 1, Kramer's V statistics to Table 2 to indicate the strength of the association between categorical variables in addition to statistical significance. We also included the cutpoints of the leadership behavior score for each tertile in the headings.

1. Authors note different findings with gender and its impact in the workplace and leadership roles. Authors may want to consider asking about the gender of the leader who is being rated in future surveys to help address some of the issues (indirectly) they note in the Discussion section.

Thank you for this input. We shared it with our PWAC team and will discuss its feasibility in the future. Also note that the congruence between gender of the leader, the person evaluating the leader is an interesting aspect we have, in part, looked at before https://pubmed.ncbi.nlm.nih.gov/33560424/.

1. I was thinking there would be some mention for the "sleep" findings in the Discussion. It seemed like it was included in the survey for some other set of reasons but not given very much attention in the text of the manuscript. The work could benefit by either removing it to avoid being distracting or to embrace it by adding more explanation around the choice for inclusion and interpretation of findings.

Again, we addressed this weakness by removing the sleep variable from our analyses.

1. A strength of the study was the large sample size and institutions involved. A fixed effects model may be helpful to consider – if authors can match the respondent to institution. There are likely some systematic differences in burnout and fulfillment by study site.

This is a great point also brought up by the first reviewer in the discussion of academic vs non-academic settings. Unfortunately, we do not have an identifier for the institutions in our dataset due to confidentiality issues.

We thank you very much for your constructive comments and helpful suggestions, which have substantially improved the manuscript.

VERSION 2 – REVIEW

REVIEWER	Messias, Erick UAMS Medical Center
REVIEW RETURNED	11-Mar-2022

GENERAL COMMENTS	I appreciate the authors' careful consideration and editing to each point listed in my review. It is my opinion that the authors' have successfully addressed each issue and there are no longer edits to be made at this point.

	Thanks.

REVIEWER	Mohr, David VA Boston Health Care System Jamaica Plain Campus
REVIEW RETURNED	28-Jan-2022

GENERAL COMMENTS	Authors appear to have been responsive in addressing both sets of reviewer comments. Changes and clarifications about tables, "sleep" variable, and availability of variables in the dataset were helpful as responses as well as changes made to the text. No further comments.