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.

This paper was submitted to a another journal from BMJ but declined for publication following peer review. The authors addressed the reviewers' comments and submitted the revised paper to BMJ Open. The paper was subsequently accepted for publication at BMJ Open.

(This paper received three reviews from its previous journal but only two reviewers agreed to published their review.)

ARTICLE DETAILS

TITLE (PROVISIONAL)	Prescription Medications for Sleep Disturbances Among Midlife Women During Two Years of Follow-up: A SWAN Retrospective Cohort Study
AUTHORS	Solomon, Daniel; Ruppert, Kristine; Habel, Laurel; Finkelstein, Joel; Lian, Pam; Joffe, Hadine; Kravitz, Howard M.

REVIEWER	Sultana, Janet University of Messina
REVIEW RETURNED	03-Aug-2020

GENERAL COMMENTS	Dear editor, Thank you for the opportunity to revise this manuscript. Overall the paper is well-written. The topic, the effectiveness of sleeping medications, is of interest. However, I have several reservations, which are listed below, and would suggest that the paper is significantly revised prior to publication. I am not convinced that the study design and data used are appropriate to answer the research question concerning drug effectiveness but would be happy to read a revised version. Introduction - It would be useful for the authors to comment briefly on whether there are any observational studies on the effectiveness of sleep medications in the same way that the mentioned the interventional studies. This would provide a better context for the study. Methods - Concerning the use of sleep medications at baseline, were the women specifically asked about medications related to the inclusion/exclusion criteria or were they guided in some way when answering? I ask this bearing in mind that some persons may not know what their medication is actually used for. - Regarding the exposure, I would say that the lack of information on the frequency of drug use is a significant limitation, as patients may opt to purchase these drugs but not take them. - Given the observational nature of the study, I would agree with the intention to treat approach. However, I wonder whether an ITT approach at 1 year and 2 years is justified given the lack of information on duration of treatment (even a proxy duration, such as number of days' supply in prescription) and given that sleeping medications are likely to be used not chronically but on a pro re nata basis. I feel this is a significant limitation that makes the results very

difficult to interpret. I would suggest using a shorter period such as 6 months.
- Concerning covariates, was any other information collected regarding menopause, other than status? Sleep problems may be related to menopausal symptoms so their resolution or otherwise may be related to the resolution of menopausal symptoms, such as with the use of oestrogen replacement therapy etc. Also, if I understand correctly, the use of other medications is not used in the propensity score. This means that there is a lot of potential for residual confounding, e.g. if medications which potentially impact on sleep disturbance, such as antidepressants, analgesics etc. were initiated. This is a significant limitation, if so.
- Statistical analysis: Is there any information on propensity score caliper? Can the authors take prescriber or health care plan into account in any way? These may be instrumental variables that require adjustment.
- In the methods, the sequence of selection is not that clear to me. For example, the mean time between the self-reported sleep disturbance and the initiation of sleep medications is not clear. This may potentially be related to the efficacy of the sleep medications. It is also not clear when the post-initiation sleep disturbance evaluation was carried out with respect to when treatment was initiated. This is of importance as it may be associated with poor quality recall if the evaluation was carried out long after the treatment was taken. In addition, a better description of the pre- and post-evaluations are needed: the post-initiation evaluations should be carried out within a time frame where drug effectiveness can be seen, i.e. not too close to initiation and not too distant. For non-users, it should be ascertained that the second/third evaluation has a temporal distance from their first one which is similar to that of the matched drug users. This would allow the investigators to account in some way for the sleep disturbance resolving itself spontaneously.

Results
- In table 1, I would add further information on the type of medical insurance (Medicaid/Medicare, private insurance + Medicare, private insurance etc.).
- The medication users vs. non-users are very balanced with regards to their severity of symptoms. This makes me wonder whether the non-users did not use sleep medications. Are there any identifiable contraindications? Issues relating to healthcare plan? And so on.
- In my opinion, there is missing information on the start of treatment with respect to the first and second/third evaluation in drug users; for non-users this would be the distance between the first and second/third evaluation. In a trial, these evaluations would be carried out at the same time for all patients. In an observational study, this may not be the case, but it could still have an impact on the reliability and interpretation of results.
Supplementary material
- Supplementary table 1 is very helpful to clarify the study design; concerning the assignment procedure, I would argue that observational corollary for the assignment procedure (perhaps “treatment assignment procedure” would be clearer) would be “based on clinical evaluation during routine medical visits”, or variants of this. Why do the authors suggest is should be “Self-report of using sleep medications”?

Miscellaneous
- There are some typos such as “prescriptioin”

	Conclusion - I would say that the study design and data considered do not permit such strong conclusions to be made.

REVIEWER	Dallas Seitz Cumming School of Medicine, University of Calgary
REVIEW RETURNED	03-Aug-2020

GENERAL COMMENTS	This is an observational study examining the associations between new use of sleep promoting medications and changes in self-reported quality for middle aged women using secondary analyses of a longitudinal cohort study. Overall the authors conclude that new initiation of sleep promoting medications was not associated with significant changes in sleep when compared to similar women who did not use these medications. Although this study does provide some new information about the real-world evidence for sleep medications there are limited conclusions about whether these medications are effective or safe in the population studied given the limitations of the study design. Comments: 1.) Abstract: The study design is a retrospective cohort study not a comparative effectiveness study design. 2.) A major limitation is that the baseline sleep disturbance scores are measured at some point after medications were administered. The relatively similar baseline scores between the exposure groups may indicate that the medications actually were effective as the scores in the group who are exposed may have actually been higher prior to treatment. The secondary analysis presented by the authors where the propensity score is derived from the visit preceding exposure to medications suggests that this might be the case. This needs to be addressed in a more substantial fashion in the manuscript. Some additional information about the duration of time between initiating treatment and the index date assessment should also be provided to help readers understand the duration of treatment that occurred prior to baseline. 3.) I think that some additional information should be provided about over the counter medications, even if these are not captured accurately in the data this information should be presented or at least discussed in terms of how frequently these medications are used in the population. 4.) The medication exposure does not include information about chronic as needed use of sleeping medications. It is not uncommon for people to have as needed sleep promoting medications which are prescribed for long periods of time even though the frequency with which they are used is low. 5.) The lack of information of about the duration of exposure is potentially problematic given that the outcome is assessment occurs 1 year after exposure information and the time-varying nature of the exposure is not accounted for in the analysis. 6.) How were individuals who were exposed to multiple medications accounted for in the analysis? 7.) An analysis of persistent users (reporting use at baseline and year 1) should be presented to help better understand if chronic use is or isn't effective. 8.) It is unclear how individuals who were non-users at index were treated if they were found to be users at either year 1 or year 2. 9.) Does the outcome measure correspond to polysomnography related outcomes? The scale employed in this survey may be neither sensitive nor specific and with only 5 measurement intervals

	it may not be sensitive to change. There is some information presented in the discussion that the measure used in this study is valid although the details of the reliability and validity of the outcome measure should be provided in greater detail in the manuscript. 10. The analysis that only includes women with more severe baseline sleep disturbances may be problematic as users were not matched on baseline severity so the final group analysed would not retain the original matched pairs. This analysis should be restricted to pairs where both exposed/unexposed pairs or triplets are concordant for the severity of their sleep disturbance. 11.) While the use of propensity scores is a strength, by only matching on the propensity score it makes it challenging to explore effect modification as the response (or non-response to medications) could differ in different age groups, those with sleep disorders, etc. where matching on these variables would allow for assessment of differential effects in subgroups.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 2

Comments: Thank you for the opportunity to revise this manuscript. Overall the paper is well-written. The topic, the effectiveness of sleeping medications, is of interest. However, I have several reservations, which are listed below, and would suggest that the paper is significantly revised prior to publication. I am not convinced that the study design and data used are appropriate to answer the research question concerning drug effectiveness but would be happy to read a revised version.
Thanks for the input.

Introduction
- It would be useful for the authors to comment briefly on whether there are any observational studies on the effectiveness of sleep medications in the same way that the mentioned the interventional studies. This would provide a better context for the study.
We were not able to identify any such studies and have added the following comments to page 4:
"Thus, effectiveness data would be useful for patients and clinicians if it included sleep medications used over several months in populations of typical patients with sleep disturbances; we found no such studies in the literature."

Methods
- Concerning the use of sleep medications at baseline, were the women specifically asked about medications related to the inclusion/exclusion criteria or were they guided in some way when answering? I ask this bearing in mind that some persons may not know what their medication is actually used for.
We did not prompt women for sleep medications. Instead, they were asked to bring in all of their medications. We have discussed this issue on page 5 of the revised manuscript: "Women were not prompted specifically about sleep medications."

- Regarding the exposure, I would say that the lack of information on the frequency of drug use is a significant limitation, as patients may opt to purchase these drugs but not take them.
We agree and have included information about the limitation in the revised manuscript (see page 8):

“Medication use was collected only at annual or biennial study visits, and there may have been intermittent use or non-adherence between visits. This is a limitation of many retrospective cohort medication analyses and limits the inferences that can be drawn.”

- Given the observational nature of the study, I would agree with the intention to treat approach. However, I wonder whether an ITT approach at 1 year and 2 years is justified given the lack of information on duration of treatment (even a proxy duration, such as number of days’ supply in prescription) and given that sleeping medications are likely to be used not chronically but on a pro re nata basis. I feel this is a significant limitation that makes the results very difficult to interpret. I would suggest using a shorter period such as 6 months.

We agree with the reviewer that shorter term follow-up data might be helpful. However, in regards to the time frame for this analysis, participants in SWAN had study visits only once per year. While occasionally, two visits occurred within a 12 month period, they were never at 6 month intervals. Thus, we do not have data at shorter intervals to analyze (see page 8): “We do not have measures of daytime consequences in this dataset. It is also possible that sleep medications may have helped in the short-term, i.e., at 8 or 12 weeks. Women only reported medication use and sleep disturbances at annual visits and thus interim outcomes (i.e., at six month intervals) are not available for analysis.”

- Concerning covariates, was any other information collected regarding menopause, other than status? Sleep problems may be related to menopausal symptoms so their resolution or otherwise may be related to the resolution of menopausal symptoms, such as with the use of oestrogen replacement therapy etc. Also, if I understand correctly, the use of other medications is not used in the propensity score. This means that there is a lot of potential for residual confounding, e.g. if medications which potentially impact on sleep disturbance, such as antidepressants, analgesics etc. were initiated. This is a significant limitation, if so.

We have added antidepressants and analgesics to Table 1.

- Statistical analysis: Is there any information on propensity score caliper? Can the authors take prescriber or health care plan into account in any way? These may be instrumental variables that require adjustment.

We do not have information on the prescriber and very limited information on health insurance.

- In the methods, the sequence of selection is not that clear to me. For example, the mean time between the self-reported sleep disturbance and the initiation of sleep medications is not clear. This may potentially be related to the efficacy of the sleep medications. It is also not clear when the post-initiation sleep disturbance evaluation was carried out with respect to when treatment was initiated. This is of importance as it may be associated with poor quality recall if the evaluation was carried out long after the treatment was taken. In addition, a better description of the pre- and post-evaluations are needed: the post-initiation evaluations should be carried out within a time frame where drug effectiveness can be seen, i.e. not too close to initiation and not too distant. For non-users, it should be ascertained that the second/third evaluation has a temporal distance from their first one which is similar to that of the matched drug users. This would allow the investigators to account in some way for the sleep disturbance resolving itself spontaneously.

The reviewer asks an important question which was not clear enough in our original manuscript. Because the timing of sleep medication initiation was not precise (did it start a week, a month or longer before the baseline visit?), we defined the baseline sleep disturbances and covariates at the first visit when a sleep medication was reported (for sleep medication users). We also conducted a secondary analysis that used the visit before to define patient characteristics (see page 6): “Other secondary analyses used the visit before sleep medication initiation to define the baseline patient characteristics to calculate the propensity score; this analysis allows us to assess the sensitivity of the results to the timing of variable measurement.”

Both analyses gave very similar results (see Supplementary Table 3).

Results
- In table 1, I would add further information on the type of medical insurance (Medicaid/Medicare, private insurance + Medicare, private insurance etc.).
See above comment.

- The medication users vs. non-users are very balanced with regards to their severity of symptoms. This makes me wonder whether the non-users did not use sleep medications. Are there any identifiable contraindications? Issues relating to healthcare plan? And so on.
Many patients have sleep disturbances but decide not to use pharmacologic interventions. There are some relative contraindications, that might include pre-existing comorbidities. We have listed baseline comorbidities in Table 1 and found no differences between the women who did and did not use sleep medications.

- In my opinion, there is missing information on the start of treatment with respect to the first and second/third evaluation in drug users; for non-users this would be the distance between the first and second/third evaluation. In a trial, these evaluations would be carried out at the same time for all patients. In an observational study, this may not be the case, but it could still have an impact on the reliability and interpretation of results.
All subjects are queried at all study visits for all prescription medications. The users and non-users of medications for sleep are queried identically. It is true, as noted above, that the exact start and stop dates of medication use are not precisely known.

Supplementary material - Supplementary table 1 is very helpful to clarify the study design; concerning the assignment procedure, I would argue that observational corollary for the assignment procedure (perhaps “treatment assignment procedure” would be clearer) would be “based on clinical evaluation during routine medical visits”, or variants of this. Why do the authors suggest is should be “Self-report of using sleep medications”?
We have followed the reviewer’s suggestions and made changes in Supplementary Table 1.

Miscellaneous
- There are some typos such as “prescriptioin”
This has been corrected.

Conclusion - I would say that the study design and data considered do not permit such strong conclusions to be made.
We have tempered our conclusions on page 9: “The current observational study does not support long-term use of sleep medications, as there were no self-reported differences at one- or two-years of follow-up comparing sleep medication users to non-users. While we used rigorous pharmaco-epidemiologic methods, the findings reported herein are based on a non-randomized observational dataset and must be seen in that light.”

Reviewer: 3

Comments:
This is an observational study examining the associations between new use of sleep promoting medications and changes in self-reported quality for middle aged women using secondary analyses of a longitudinal cohort study. Overall the authors conclude that new initiation of sleep promoting medications was not associated with significant changes in sleep when compared to similar women who did not use these medications. Although this study does provide some new information about the

real-world evidence for sleep medications there are limited conclusions about whether these medications are effective or safe in the population studied given the limitations of the study design. We have tempered our conclusions as noted above.

Comments:
1.) Abstract: The study design is a retrospective cohort study not a comparative effectiveness study design.
This has been revised per the reviewer's suggestion.

2.) A major limitation is that the baseline sleep disturbance scores are measured at some point after medications were administered. The relatively similar baseline scores between the exposure groups may indicate that the medications actually were effective as the scores in the group who are exposed may have actually been higher prior to treatment. The secondary analysis presented by the authors where the propensity score is derived from the visit preceding exposure to medications suggests that this might be the case. This needs to be addressed in a more substantial fashion in the manuscript. Some additional information about the duration of time between initiating treatment and the index date assessment should also be provided to help readers understand the duration of treatment that occurred prior to baseline.
The reviewer highlights the results from a secondary analysis of a secondary outcome that showed a subtle difference in baseline to year 2 results for early morning awakenings (see Supplementary Table 3). While we do not disagree with the reviewer, we are hesitant to over-interpret this finding. If one inspects Supplementary Table 3, the differences observed are very subtle.

3.) I think that some additional information should be provided about over the counter medications, even if these are not captured accurately in the data this information should be presented or at least discussed in terms of how frequently these medications are used in the population.
We have inconsistently collected information on over the counter (OTC) medications. At the urging of this reviewer, we examined reports at the baseline visit of OTC sleep medications including melatonin, diphenhydramine, and doxylamine. Among the sleep medication users, 37 (15.6%) reported OTC sleep medication use; 41 (9.2%) of those who did not use sleep medications reported OTC use. This information was added to the discussion (see page 9, paragraph 1): "We know that 11% of the women in this study reported use of an over-the-counter hypnotic at the baseline visit; slightly more women in the user group reported such use compared with the non-user group."

4.) The medication exposure does not include information about chronic as needed use of sleeping medications. It is not uncommon for people to have as needed sleep promoting medications which are prescribed for long periods of time even though the frequency with which they are used is low.
We agree that this is a limitation of the study dataset. We have discussed this limitation in the revised manuscript on page 8: "Women only reported medication use and sleep disturbances at annual visits and thus interim outcomes (i.e., at six month intervals) and intermittent medication use are not available for analysis."

5.) The lack of information of about the duration of exposure is potentially problematic given that the outcome is assessment occurs 1 year after exposure information and the time-varying nature of the exposure is not accounted for in the analysis.
We agree and have described this limitation on page 8: "Medication use was collected only at annual or biennial study visits, and there may have been intermittent use or non-adherence between visits."

6.) How were individuals who were exposed to multiple medications accounted for in the analysis?
Women who used multiple sleep medications were included in the sleep medication user category, without distinction.

7.) An analysis of persistent users (reporting use at baseline and year 1) should be presented to help better understand if chronic use is or isn't effective.
All analyses accounted for persistence as evidenced by smaller N's at year 2.

8.) It is unclear how individuals who were non-users at index were treated if they were found to be users at either year 1 or year 2.
Non-users were never users. This point has been clarified in the revised manuscript (see page 5, paragraph 4): "Non-users were not included if they became users at a later visit."

9.) Does the outcome measure correspond to polysomnography related outcomes? The scale employed in this survey may be neither sensitive nor specific and with only 5 measurement intervals it may not be sensitive to change. There is some information presented in the discussion that the measure used in this study is valid although the details of the reliability and validity of the outcome measure should be provided in greater detail in the manuscript.
The sleep disturbance questions asked in SWAN are nearly identical to the Women's Health Initiative Insomnia Rating Scale (WHIIRS). We were unable to find any publication in which the WHIIRS was compared directly to polysomnogram measures. However, the developers of the WHIIRS (new citation #19) have compared this instrument with actigraphy. They found significant correlations in the expected direction for trouble falling asleep with wake after sleep onset (WASO) and sleep latency, waking several times at night with WASO and sleep efficiency, and waking earlier than planned with sleep efficiency. In other words, the objective actigraphy measures correlated most highly with the WHIIRS items that were intended to tap into the same aspect of the insomnia construct. Levine et al (new citation #18) also validated the 5-item WHIIRS (which included our 3 sleep items) in a large clinical trial involving postmenopausal women.

10. The analysis that only includes women with more severe baseline sleep disturbances may be problematic as users were not matched on baseline severity so the final group analysed would not retain the original matched pairs. This analysis should be restricted to pairs where both exposed/unexposed pairs or triplets are concordant for the severity of their sleep disturbance.
The current sensitivity analysis focusing on women with more severe baseline sleep disturbances retains the match. This has been clarified on page 6, paragraph 4:
"Such analyses retained the propensity score match."

11.) While the use of propensity scores is a strength, by only matching on the propensity score it makes it challenging to explore effect modification as the response (or non-response to medications) could differ in different age groups, those with sleep disorders, etc. where matching on these variables would allow for assessment of differential effects in subgroups.
This is an interesting comment and we agree that there may be interesting subgroups. However, we felt that it was more important to look across the entire population as a first step in the analysis. Finding no differences, we did not deem it worthwhile to look at subgroups.

VERSION 2 – REVIEW

REVIEWER	Janet Sultana University of Exeter, UK
REVIEW RETURNED	02-Nov-2020

GENERAL COMMENTS	Dear authors and editor, Thank you for forwarding this paper. As I had already reviewed it in detail for BMJ before it was transferred to BMJ Open, there is not very much to add in terms of comments. Overall, I think the authors

	addressed several issues and the paper is much clearer now. My main reservation is that, in my opinion, the study does not really provide results on effectiveness per se. What I mean is that the study design does not permit inference of effectiveness. To my mind is it studies such as this that do: doi: 10.1002/edm2.103 . I think the strength of the study is the patient reported outcomes and how these change over the study period. I would focus on this kind of discourse rather than on "real-world effectiveness", as in my opinion the latter is misleading and is not reflected in the paper. Perhaps this is something that should be discussed with the editor.

REVIEWER	Wessel van Leeuwen Stockholm University, Sweden
REVIEW RETURNED	25-Nov-2020

GENERAL COMMENTS	Ethics approval statement is missing. For the rest, the manuscript has obviously already gone through a number of peer reviews - the critics of which have been met in a satisfactory way. The limitations that are obviously intrinsic to the used methodology have been clearly put forward by the authors.

REVIEWER	Nicholas Vozoris University of Toronto, Canada
REVIEW RETURNED	26-Nov-2020

GENERAL COMMENTS	This manuscript covers an important topic, on which there is little published research, and contains clinically meaningful messaging to health care providers. My comments/suggestions are as follows: 1. Throughout the manuscript, the authors have written that the Z-drugs (zaleplon, zolpidem and eszopiclone) are "non-benzodiazepine hypnotics". This is incorrect. They are selective benzodiazepine receptor agonists, having selective affinity for the GABA-A alpha-1 receptor, whereas drugs like lorazepam, clonazepam, etc. are non-selective benzodiazepine receptor agonists. For further explanation about this, refer to: Vozoris NT. Benzodiazepine and opioid co-usage in the US population, 1999-2014: an exploratory analysis. Sleep. 2019;42(4):zsy264. Please correct the classification naming of Z-drugs throughout the manuscript. Along these same lines, the sensitivity analysis in Table 4 should be renamed non-selective benzodiazepine receptor agonists versus no sedative use. 2. How did non-users enter the cohort and what was their index date? This should be clearly specified in the methods section. 3. Variables included in the propensity score model were evaluated at the index date and outcomes were evaluated far along off, at one and two years out. Some variables included the propensity score model (e.g., depression, anxiety, pain, comorbidity status, etc.) may change in an individual over time, which can then have an influence on the occurrence of outcomes. How was this addressed? If not addressed, identify as a limitation. 4. Somewhere in the manuscript (perhaps best in an Appendix) describe the variables depression score, anxiety score, pain score, SF-36 Mental and SF-36 Physical that were included in the propensity score model. Many readers may be unfamiliar with these

scoring systems.

5. In Tables 2 & 3, the authors write that about 70-80% of both exposed and control groups had at least one of the three insomnia complaints at 3/week. However, shouldn't that number for both groups be 100%, because when describing the study population on page 6 the authors write: "we required all women to have reported during SWAN follow-up a sleep disturbance on at least 3 nights per week during a two-week interval." How is this discrepancy explained? This is an important point that must be adequately addressed.

6. Consider undertaking a sensitivity analysis of selective benzodiazepine receptor agonists (i.e., Z-drugs) versus no sedative use, since the former drugs are specifically marketed as sleep aids (whereas non-selective benzodiazepine receptor agonists have other recognized indications, besides as sleep aids) and are popularly prescribed.

REVIEWER	Chaudhary NS UAB, USA
REVIEW RETURNED	05-Dec-2020

GENERAL COMMENTS

The authors have presented an interesting paper focused on determining the long-term effectiveness of sleep medications among peri-menopausal women. Using propensity score method and based on the “target trial emulation” concept, the authors found that those who were prescribed sleep medications did not show improvement in their sleep disturbances compared to those who were not. The authors concluded that the long-term use of medication should be re-examined. This is an exciting work and excellent approach to leverage the use of population-based study. The question raised by the authors is important. Below are my comments:
1. The main concern/missing component about the paper is the identification of exposure groups being effectiveness study. The sleep disturbances seem similar between medication users and non-users based on Table 2 which makes it important to know how these factors differ between these groups or why one group seems to be prescribed on medications while others were not. For that, it will be helpful if the authors can provide the comparison of characteristics between sleep medication users (n=260) and non-users (n=1268) is required. This information will also help to know how important propensity matching for the groups was.
2. Considering this is a study based on peri-menopausal women, it does not seem authors have accounted for the effect of estrogen replacement therapy. The effect of estrogen replacement therapy on sleep patterns is well-studied and needs to be adjusted for or accounted for in the propensity score.
3. Propensity Score Models:
a. The propensity models should be as non-parsimonious as possible for better balance and adding additional covariates will be helpful. For e.g. being a multi-site study, there is a possibility that prescription patterns may vary by site. Did authors include site in the model?
b. Can the authors provide distribution of the propensity score overall and by groups? The concern is if predicted probabilities are discrete, then it may induce pseudo-balance.
c. The authors have not clearly specified caliper width. Is it 0.2 of standard deviation of logit of propensity score?

	d. Considering for N=238 there are only 447 participants that matched (and not 476), the authors should clarify that not all were matched in the ratio of 2:1 or whether few participants serve as match for more than one medication users. 4. Analysis: The authors have not commented on proportional odds assumption. 5. It is bit confusing to understand how sleep medication user group is different from prevalent users group that was excluded. Do you have information on when sleep medication was initiated for sleep medication user group? Was index visit different for each participant? 6. “Non-users were not included if they became users at a later visit” – it seems this group is not represented in the inclusion/exclusion figure. 7. Have authors considered the association with primary outcome stratified by propensity score tertiles or quartiles? I am wondering if it may address some bias related to non-adherence or frequent dosing. For e.g. those who are highest tertiles/quartile can likely be those who need to be sleep medication and may be adherent to medication. 8. While I understand the objective of the paper, with no or limited information on adherence or change in prescription patterns, the authors can state that they are studying long-term effectiveness but stating that they are studying long-term medication use is bit concerning.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1

1. Thank you for forwarding this paper. As I had already reviewed it in detail for BMJ before it was transferred to BMJ Open, there is not very much to add in terms of comments. Overall, I think the authors addressed several issues and the paper is much clearer now.

Reply: Thanks.

2. My main reservation is that, in my opinion, the study does not really provide results on effectiveness per se. What I mean is that the study design does not permit inference of effectiveness. To my mind is it studies such as this that do: doi: 10.1002/edm2.103 . I think the strength of the study is the patient reported outcomes and how these change over the study period. I would focus on this kind of discourse rather on the “real-world effectiveness”.

Reply: Whereas the authors’ understanding is that “effectiveness” typically describes the potential benefits of an intervention in typical care (and “efficacy” refers to trial benefits), we are aware that there are different definitions used in the literature. In deference to the reviewer, we have acknowledged the limitations of our methods, as described in the discussion. Furthermore, we have deleted “effectiveness” and inserted the term “retrospective cohort study” in the revised manuscript. We will defer to the Editor if s/he has another suggestion.

Reviewer: 2

1. Ethics approval is missing. For the rest, the manuscript has obviously already gone through a number of peer reviews - the critics of which have been met in a satisfactory way. The limitations that are obviously intrinsic to the used methodology have been clearly put forward by the authors.

Reply: We appreciate the reviewer’s comments and have added the details about ethics approval (see page 15) and have mentioned the patient consent process on page 5 in the section “Patient and public involvement.”

Reviewer: 3

This manuscript covers an important topic, on which there is little published research, and contains clinically meaningful messaging to health care providers.

Reply: Thanks.

1. Throughout the manuscript, the authors have written that the Z-drugs are “non-benzo hypnotics.” This is incorrect. They are selective benzo receptor agonists, having selective affinity for specific receptors. Please correct the classification naming of Z-drugs throughout the manuscript. Along these same lines, the sensitivity analysis in Table 4 should be renamed non-selective benzodiazepine receptor agonists versus no sedative use.

Reply: We have corrected this issue throughout the manuscript selective benzodiazepine receptor agonists (Z-drugs).” We have changed the label in Table 4. We thank the reviewer for their input.

2. How did non-users enter the cohort and what was their index date?

Reply: Non-users never reported use of a prescription sleep medication during follow-up. They entered the study (index date) at visits matched in frequency distribution with the sleep medication user. We have clarified this further in the revised manuscript (see page 5, paragraph 4):

“Non-users were not included if they became users at a later visit. Non-users entered the study (index date) at visits matched in frequency distribution with the sleep medication user.”

3. Variables included in the propensity score model were evaluated at the index date and outcomes were evaluated far along off, at one and two years out. Some variables included the propensity score model (e.g., depression, anxiety, pain, comorbidity status, etc.) may change in an individual over time, which can then have an influence on the occurrence of outcomes. How was this addressed? If not addressed, identify as a limitation.

Reply: We recognize that covariates change during follow-up. However, covariates were well balanced at baseline. Just as with a randomized controlled trial, changes in covariates could be related to exposure (or non-exposure) to sleep medications. Thus, we chose not to update covariates and have added this to the potential limitations (see page 8, paragraph 3):

“We did not update covariates in the two-year analysis.”

4. Somewhere in the manuscript describe the variables, such as depression score, anxiety score, pain score, SF-36 mental, SF-36 physical that were included in the propensity score model. Many readers may be unfamiliar with these scoring systems.

Reply: We have added description of these scores to the footnote of Table 1.

5. In Tables 2 & 3, the authors write that about 70-80% of both exposed and control groups had at least one of the three insomnia complaints at 3/week. However, shouldn't that number for both groups be 100%, because when describing the study population on page 6 the authors write: "we required all women to have reported during SWAN follow-up a sleep disturbance on at least 3 nights per week during a two-week interval." How is this discrepancy explained? This is an important point to address.

Reply: 100% of women included reported a sleep disturbance at some point during follow-up. At baseline, 72-77% reported sleep disturbance. We have clarified this point in the revised manuscript at the bottom of page 6:

“100% of women included reported a sleep disturbance at some point during follow-up. At baseline, 72-77% reported sleep disturbance.“

6. Consider undertaking a sensitivity analysis of selective benzo receptor agonists (Z-drugs) versus no medication use.

Reply: We have produced such an analysis (see Table below). The results are similar to the current sensitivity analysis shown in Table 4 (no difference between medication users and non-users) and thus was not added to the manuscript.

Table: Z-drug Users Versus No Medication Use

	Year 0 vs Year 1	
	Estimate	P-value
Difficulty initiating sleep	0.08	0.75
Waking Frequently	0.05	0.82
Early morning awakening	-0.1	0.69

Reviewer: 4

The authors have presented an interesting paper focused on determining the long-term effectiveness of sleep medications among peri-menopausal women. Using propensity score method and based on the "target trial emulation" concept, the authors found that those who were prescribed sleep medications did not show improvement in their sleep disturbances compared to those who were not. The authors concluded that the long-term use of medication should be re-examined. This is an exciting work and excellent approach to leverage the use of population-based study. The question raised by the authors is important.

Reply: Thanks.

1. The main concern/missing component about the paper is the identification of exposure groups being effectiveness study. The sleep disturbances seem similar between medication users and non-users based on Table 2 which makes it important to know how these factors differ between these groups or why one group seems to be prescribed on medications while others were not. It will be helpful if the authors can provide a comparison of the women who were sleep medication users with the non-users. This information will also help to know how important propensity matching for the groups was.

Reply: We have added a supplemental table with this information (see new Supplementary Table 2). There were many variables demonstrating imbalance. This is the rationale for the use of a propensity score matched analysis, allowing for a more valid set of comparisons.

2. Considering this is a study based on peri-menopausal women, it does not seem authors have accounted for the effect of estrogen replacement therapy the effect of estrogen replacement therapy. The effect of estrogen replacement therapy on sleep patterns is well- studied and needs to be adjusted for.

Reply: See below.

3. Propensity Score Models:

a. The propensity models should be as non-parsimonious as possible for better balance and adding additional covariates will be helpful. For e.g. being a multi-site study, there is a possibility that prescription patterns may vary by site. Did authors include site in the model?

Reply: We added site and estrogen replacement therapy into the models (see Table below). The results were very similar to the original results. This table has been added as a supplement (see Supplemental Table 6).

Table: Original models additionally adjusted for site and estrogen use

	Year 0 vs Year 1		Year 0 vs Year 2		Year 1 vs Year 2	
	Estimate	P-value	Estimate	P-value	Estimate	P-value

Difficulty initiating sleep	-0.15	0.3	-0.18	0.04	-0.24	0.21
Waking Frequently	-0.03	0.83	0.11	0.22	0.21	0.02
Early morning awakening	-0.19	0.22	-0.12	0.17	-0.07	0.45
The significant p-value in Year 0 vs Year 2 shows that the non users decreased by .12 and the med users increased by .06. The significant p-value in Year 1 vs Year 2 shows that non users decreased by .004 and the med users decreased by .22.						

b. Can the authors provide distribution of the propensity score overall and by group?

Reply: We show below the PS distributions below for the overall study population and by group. These graphs demonstrate good balance.

Overall:

By Group:

c. The authors have not clearly specified caliper width. Is it 0.2 of a standard deviation of the logit of propensity score?

Reply: Yes, this is correct and now stated as noted above on page 6, paragraph 2 (last sentence of “Statistical analyses”) of the revised manuscript.

d. Considering for N=238 there are only 447 participants that matched and not 476, the authors should clarify.

Reply: The reviewer is correct that we attempted to match 2:1 but some participants could not be matched. We have clarified this issue in the revised manuscript (“Results” 1st paragraph; see page 6, last paragraph):

“Thus, we propensity matched the 238, attempting, attempting to find 2 non-users for each user; we were able to match 447 women who never initiated a sleep medication during study follow-up.”

4. Analysis: The authors have not commented on proportional odds assumption.

Reply: All proportional odds assumptions were met. This has now been stated on page 7, paragraph 4 (“Results” end of last paragraph).

5. It is bit confusing to understand how sleep medication user group is different from prevalent users group that was excluded. Do you have information on when sleep medication was initiated for sleep medication user group? Was index visit different for each participant?

Reply: The index visit was different for each participant who reported starting a sleep medication during SWAN follow-up. We show below the distribution of SWAN visits considered the index visit for this study. We matched index visits in the two groups based on frequency distribution as demonstrated in the Table below.

Table: Distribution of Index Visits and Sleep Medication Users

	Total	NO SLEEP MEDICATIONS	SLEEP MEDICATION USERS	P-value

	N=685	n=447	n=238	
Index visit				0.79
1	67 (9.8)	41 (9.2)	26 (10.9)	
2	75 (11.0)	53 (11.9)	22 (9.2)	
3	63 (9.2)	41 (9.2)	22 (9.2)	
4	62 (9.1)	43 (9.6)	19 (8.0)	
5	47 (6.9)	28 (6.3)	19 (8.0)	
6	1 (7.5)	32 (7.2)	19 (8.0)	
7	56 (8.2)	38 (8.5)	18 (7.8)	
8	44 (6.4)	30 (6.7)	14 (5.9)	
9	56 (8.2)	41 (9.2)	15 (6.3)	
10	70 (10.2)	41 (9.2)	29 (12.8)	
12	94 (13.7)	59 (13.2)	35 (14.7)	

6. “Non-users were not included if they became users at a later visit”. It seems this group is not represented in Figure 1.
Reply: We have corrected the statement in the paper to clarify that non-users were actually never users (“Exposures” end of 2nd paragraph; see page 5, paragraph 4). Thus, women who became users were included in the medication user group.
“Non-users were never users.”

7. Have authors considered the association with primary outcome stratified by propensity score quantile, e.g. those who are highest tertiles/quartile can likely be those who need to be sleep medication and may be adherent to medication.
Reply: We provide below a PS-quartile stratified analysis. There were no substantial differences between this analysis and the primary analysis.

Table: Propensity Score Quartile Stratified Analyses

	PS Quartile 1		PS Quartile 2		PS Quartile 3		PS Quartile 4	
	Estimate	P-value	Estimate	P-value	Estimate	P-value	Estimate	P-value
Difficulty initiating sleep	-0.27	0.27	0.26	0.37	-0.28	0.36	-0.30	0.34
Waking Frequently	-0.11	0.72	-0.16	0.62	0.21	0.48	-0.10	0.65
Early morning awakening	-0.48	0.11	-0.36	0.23	-0.17	0.63	-0.09	0.74

8. While I understand the objective of the paper, with no or limited information on adherence or change in prescription patterns, the authors can state that they are studying long-term effectiveness but not long-term medication use.
Reply: We have removed mention of “long-term medication use” from the revised manuscript.

VERSION 3 – REVIEW

REVIEWER	Janet Sultana University of Exeter, UK
REVIEW RETURNED	09-Jan-2021

GENERAL COMMENTS	Dear authors, Since I have reviewed this paper twice already, I do not have much to add except that it has improved significantly in clarity – well done indeed. To my mind the main value of this paper is the patient-reported outcome aspect, which is not as common as one would hope in pharmacoepidemiology. The authors also leverage the propensity score method to adjust for a large number of covariates. In my opinion, it might be to the authors' advantage to discuss the patient-reported component of the study in more detail (in the intro or discussion), highlighting its value. The only minor comment I have at this point is to consider reporting the means with 95% CI rather than the STD, because the STD is so large compared to the mean that it is difficult to interpret and is not so informative. This would also help the abstract to read better as the main results reported here are mean/STD. Also I didn't seem to find the STROBE documentation which I believe is required. In general I believe this paper is ready to be published.

REVIEWER	Dr. Nicholas Vozoris University of Toronto, Canada
REVIEW RETURNED	22-Jan-2021

GENERAL COMMENTS	The authors have addressed my comments to my satisfaction.

REVIEWER	Ninad S Chaudhary UAB, USA
REVIEW RETURNED	21-Jan-2021

GENERAL COMMENTS	The authors have provided a satisfactory response to my questions and have made the necessary changes. I do not have any further suggestions/recommendation.

VERSION 3 – AUTHOR RESPONSE

Reviewer: 1
1. Since I have reviewed this paper twice already, I do not have much to add except that it has improved significantly in clarity – well done indeed. To my mind the main value of this paper is the patient-reported outcome aspect, which is not as common as one would hope in pharmacoepidemiology. The authors also leverage the propensity score method to adjust for a large number of covariates. In my opinion, it might be to the authors' advantage to discuss the patient-reported component of the study in more detail (in the intro or discussion), highlighting its value.
Reply: We have added text to the discussion highlighting the patient-reported component of the sleep difficulty outcome (see page 8, last paragraph):
"Furthermore, the use of patient-reported sleep disturbances (or other symptoms) can be viewed as a strength, since it reflects how patients experience sleep disturbances."

2. The only minor comment I have at this point is to consider reporting the means with 95% CI rather than the STD, because the STD is so large compared to the mean that it is difficult to interpret and is not so informative. This would also help the abstract to read better as the main results reported here are mean/STD. Also I didn't seem to find the STROBE documentation which I believe is required. In

general I believe this paper is ready to be published.
Reply: We have substituted 95% CIs for standard deviations throughout the manuscript, except in Table 1.

Reviewer: 3.The authors have addressed my comments to my satisfaction.
Reply: Thanks.

Reviewer: 4 The authors have provided a satisfactory response to my questions and have made the necessary changes. I do not have any further suggestions/recommendation.
Reply: Thanks.