Response to the Stats Editor

Stats Editor Comments  
  
**C1:** I enjoyed the mix of a tutorial and an original manuscript that the authors have developed. The writing is very clear, esp the sections covering the assumptions. I would suggest moving this to a tabular presentation as it may be easier to review and would make a nice exhibit.  
  
**R1**: We thank the statistical reviewer for his/her positive comment and his/her enthusiasm about our manuscript. We agree with the proposal and we move the assumption section in a tabular presentation in the manuscript in Table/Figure XX

Response to the Reviewer: 1

**C**: Inverse Probability Treatment Weighting (IPTW) and IPTW with Regression Adjustment (IPTW- RA) go one step beyond standard adjusting and matching and can therefore contribute to reducing or removing the impact of confounders in a given analysis, ultimately aiming for causal inference. These methods have been successfully applied to a cohort of 9,244 OSA patients treated with CPAP, showing the feasibility of using them to obtain relevant scientific results. The authors are very pedagogic when explaining both methods (and the assumptions being held), and provide the code used to perform the analyses.

**R**: we thank the reviewer for his/her positive evaluation of our manuscript.

**C1**: It would be illustrative to show the results of a standard regression analysis including confounders rather than just using a crude T test. In fact, it could be argued that the same multilevel (0-4, 4-6, 6-7, 7-10) approach could be used with standard regression methods.

**R1**: This comparison was initially proposed but removed by considering the length of the manuscript. We add a sentence in the text.

**C2:** Figure 1. Why is age not directly connected to comorbidities? What is the meaning of colour boxes in the figure?

**R2**: Age is connected to comorbidities, however to make the figure easy to read we did not show all arrows. We have added a footnote specifiying this in the Figure one.

**C3**: Figure 2 is missing, and not being referred to in the text.

**R3: A REVOIR FRANCOIS**

**Figure 2 is referred in text in material and method subsection Weight estimation and balancing properties just below the second formula.**

**C4**: Figure 5 may be difficult to understand as presented. What does “change in Epworth” refer to? Please, include further details in the figure’s legend.

**R4**: à developer François : tu peux faire une proposition ?.

**C5**: The discussion could be missing a reinforcement of the message that IPTW / IPTWRA are providing results that are much stronger (in terms of causality) than standard methods, and thus closer to the truth. Unfortunately, one could misinterpret Fig 5 and assume that standard methods provide better results (bigger differences between compared groups). This kind of “wishful thinking interpretation” or bias towards positive results must be combated through pedagogy and this paper is the perfect arena to do so.

C5: Clémence ?

Minor comments:

**C6**: Introduction. Review sentence: “In this study, we propose to explain, illustrate the use of two different causal inference methods”

**R6**: Clémence ?

**C7**: Methods. The standard definition of a confounder includes being related to both the exposure and the outcome. Therefore, the sentence “all potential confounders must be accounted for (i.e. all variables that can be related with Epworth score or CPAP adherence must be considered)” is not fully accurate as it should use “and” instead of “or”.

**R7**: The sentence is now replaced by: “with Epworth score **and** CPAP adherence”.

**C8**: Table 2. Please correct the sentence “1 means that there is a statistically significant difference between the 0-4h adherence group and the 0-4h adherence group (1) for the variable in question.

**R8**: François : àà corriger

**C9**: Results. Revise the use of “fold” in the sentence “When the adherence groups are compared with an unweighted mean difference, patients in 0-4 hours adherence group, have an average Epworth score of 2 (95% bootstrap confidence intervals based on percentiles 1.7; 2.3) fold higher than patients in 7-10 hours adherence group”.

**R9**: Clémence ?

Responses to Reviewer: 2

Thank you for the opportunity to revise the manuscript "Application of inverse-probability-of-treatment weighting to estimate the effect of daytime sleepiness in obstructive sleep apnea patients" written by Bettega et al. In this study, the authors show how two causal inference methods, inverse-probability-of-treatment weighting (IPTW) and IPTW with regression adjustment (IPTW-RA), can be applied on observational data for the estimation of the effect of different ranges of CPAP adherence on daytime sleepiness measured by the change of Epworth sleepiness score (ESS) over time.

I agree with the authors that "When RCTs cannot be implemented or when real-world evidence is needed, observational studies, such as cohorts or registries, contain a wealth of data for causal inference." Importantly, this study proposed a method to consider CPAP adherence as a multilevel variable, in contrast to a binary one and illustrated the application of the IPTW method to reduce confounding bias. However, I have some concerns about this manuscript as written.

MAJOR CONCERNS:

Given that this study is presented as a methodological to improve the quality of observational studies while looking on the effect of treatment, I wonder if the authors would additionally consider the following:

**C1**: Novelty. This is not the first time different propensity score (PS) methods were offered to assess CPAP effectiveness or adherence with CPAP, e.g., <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7917762/>; <https://erj.ersjournals.com/content/44/2/405>. Specifically, subclassification using propensity score quintiles, which aimed to replicate covariate balance associated with randomized trials and, therefore, minimize selection bias and allow causal inference, has been recommended: <https://erj.ersjournals.com/content/44/2/405>. I wonder why the authors have not mentioned this in the introduction section to justify why IPTW is better than other PS approaches.

**R1**: completer la biblio de l’intro avec ces ref. SB

**C2** Given that the authors consider this study as "causal inference methods in sleep apnea," I appreciate seeing Directed Acyclic Graphs (DAG) as a part of the process, given that it is usually recommended to display assumptions about the relationship between variables. As written, I am not exactly sure how this DAG was created. Some details (maybe in the Data Supplement) could be useful. For example, I would argue that some comorbidities may influence both CPAP adherence and daytime sleepiness, while this is not an assumption in the diagram. In addition, some medications or habits (e.g., alcohol consumption) may also influence both CPAP adherence and daytime sleepiness, which was not considered in the diagram.

**R2:**

**C3**: The use of these weights allows one to estimate the average treatment effect (ATE) or the average treatment effect in the treated (ATT), depending on the question: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5157758/pdf/SIM-35-5642.pdf>. As described, I am unsure which approach and why was used, which is important to clarify. As a disclosure, I am not very good at equations; so, maybe I just have missed an explanation.

**R3**: Clémence ?

**C4**:   Usually, expert opinion, literature reviewer on plausible pathogenetic mechanisms are also recommended to include in the variable selection process in addition to the statistical approach as based on p values. As described, I am not exactly sure if this was done.

**R4**: François : developepr la méthode de sélection de variable avec expert opinion (JLP, RT)

**C5** : It has been suggested that "the popular inverse probability weighting method in causal inference is often hampered by extreme propensity scores, resulting in biased estimates and excessive variance. A common remedy is to trim patients with extreme scores (i.e., remove them from the weighted analysis) [the approach used from my understanding in the current study]. However, such methods are often sensitive to the choice of cutoff points and discard a large proportion of the sample. The implications for bias and the precision of the treatment effect estimate are unclear. These problems are mitigated by a newly developed method, the overlap weighting method. Overlap weights emphasize the target population with the most overlap in observed characteristics between treatments, by continuously down-weighting the units in the tails of the propensity score distribution" (https://academic.oup.com/aje/article/188/1/250/5090958; <https://www2.stat.duke.edu/~fl35/OW.html>). Given that this is a methodological manuscript, I wonder if authors would additionally consider the overlap weighting method in comparison to two methods they have utilized.

**R5**: François, tu peux regarder et faire une proposition ?

**C6**: Given that the ESS is not the best measure of daytime sleepiness, I wonder if the authors tested their methods on other potential measures of daytime sleepiness/tiredness available, such as: "sleepiness at the wheel," "morning tiredness," "fatigue measured by Pichot's scale" to confirm the similar results. For the ESS, the authors concluded that "the causal effect of CPAP adherence on daytime sleepiness which was mainly observed between the lower CPAP adherence group (0-4h) compared to the higher CPAP adherence group (7-10h). There are no differences by considering higher level of CPAP adherence (>4h)". This suggests from a clinical perspective that we could continue recommended adherence of at least 4 hours per night, despite some studies recommend 6 hours. This is why I wonder if the same was noted for other symptoms.

**R6**: We thank the reviewer for this comment. Indeed, it would be of interest to assess this approach on different symptoms. But as it is a methodological papers which aims to present a method, we consider that it could be complex, for clinical readers, to develop both methods and outcomes. We add the following sentence in the discussion: “Further applications would be performed to investigate the use of such methods on other symptoms”.

MINOR:   
**C7**: In the discussion, the authors mentioned that "Conversely to binary exposures, a few published studies have applied IPW to multilevel exposure, such as different levels of CPAP adherence in order to assess the marginal causal effect on an outcome." However, the references were not provided, which would be important to add.

**R7**: François, tu peux vérifier ?

**C8**: I was not able to find supplementary materials and tables, which may affect my review above. I wonder if it was missed for some reason.

**R8**: François à verifier

C9: The author suggested that ordinary regression can not be used in the abstract, and then in the introduction, they mentioned that "When a study population is large enough, propensity score based methods, such as IPTW, and multivariable regression lead to similar results." I wonder if they tried to use regressions and if the results were different?

**R9**: François reprendre la regression et modifier l’abstract