PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	A Pilot Study of A Wearable Apnea Detection Device
AUTHORS	Rodriguez-Villegas, Esther; Chen, Guangwei; Radcliffe, Jeremy; Duncan, John

VERSION 1 - REVIEW

REVIEWER	Mervyn Lyons University of Dundee
REVIEW RETURNED	01-May-2014

GENERAL COMMENTS	This is a very interesting device with good sensitivity and specificity which shows considerable potential for further development.

REVIEWER	Frederic Roche Physiology, Sleep Lab Saint Etienne University Hospital France
REVIEW RETURNED	03-May-2014

GENERAL COMMENTS	The number of patients on which the analysis is done in my opinion is too small to draw definitive conclusions on the diagnostic value field situation. A real sample size calculation must be performed. It is an interesting preliminary study testing the diagnostic value of a new system of screening for the presence of apnea / hypopnea sleep from the noise tracheal recording. Results on a polygraph synchronous interactive analysis seem interesting analysis "epoch by epoch". In particular, the sensitivity seems excellent for severe forms of OSAS. There remain important issues and limitations in this study. The number of patients on which the analysis is done in my opinion is too small to draw definitive conclusions on the diagnostic value field situation. A real sample size calculation must be performed and probably more subjects would be included in this study. The somnoscreen is recognized for its automatic very low sensitivity analysis and comparison with this software can only show strong limitations of this somnocreen system. We must insist on comparing the new system with oximetry that remains the main screening tool used yet in routine clinical setting How to improve the detection of hypopneas? I confess also do not believe too much in the future use of this device for the detection of apnea in epileptic patients. I will add in the "study limitations" that this criterion (tracheal sounds

	and apnea index rated) does not allow to assess the hypoxic load or autonomic activation and therefore impact the cardiovascular or stroke associated with OSA syndrome.

REVIEWER	Matthew Strand National Jewish Health, USA
REVIEW RETURNED	27-Jun-2014

GENERAL COMMENTS	Very interesting paper and analysis. The correct statistical methods just need to be applied (if applicable) and then described more fully in the manuscript. **Statistical review** In the original review of the article (BEFORE revision), a reviewer made important comments in Major point 2 (It begins, "The authors need to be much clearer…"). Although I do not have the original manuscript, it appears to me that the current version does not address the points brought up here, particularly parts b and d, even though someone later had said that statistical issues had been addressed – I do not see that to be the case, especially regarding point 2. I will comment on 2b and 2d separately, below: 2b: The use of the tests to identify 'the truth' blurs the line between the two and will thus bias performance statistics. Thus the second approach is more sound due to the fact that 'truth' and 'test' are more clearly separated, even if there could be some error in the clinicians' judgments. Therefore, I would consider the 2nd approach to be the primary analysis despite possible limitations. On a related note, definitions in the Data Analysis section (listed as 'a)' through 'h)' use the word 'true' and 'false'. This becomes confusing if they are predictions based on tests. The authors might consider adjusting terminology. 2d: Before reviewing the comments I had also thought that there was a lack of detail regarding the repeated measures. This might impact the performance statistics but will certainly affect the confidence intervals. This brings up an even greater issue: how were the performance statistics calculated? If they were based on a longitudinal logistic regression model then the appropriate inference could be performed. Otherwise, the correlation within individuals needs to be accounted for in some fashion, particularly when computing the confidence intervals. The data cannot simply be pooled and treated as independent observations when conducting inference. I'm not sure if that was done or not, but there is no information about it that I could find in the article. Thus, the recommendations here are to (1) properly account for repeated measures in the data when determining performance statistics and related confidence intervals, and (2) describe in the Data Analysis section how this was done. Note that fitting a longitudinal logistic regression model (e.g., employing GEE

	for a generalized linear model or using a generalized linear mixed model with pseudo-likelihood methods) will give you the ability to handle serial correlation within subjects; such models will also allow you to enter covariates into the model. Minor comment: in the Abstract, please change “CI” to “95% CI”. In Table 2, change “(95 CI)” to “(95% CI)”.

VERSION 1 – AUTHOR RESPONSE

1. Please use a title that frames the research question and the study design, rather than a headline

We have changed the title of the manuscript to "A Pilot Study of A Wearable Apnea Detection Device"

2. This is a very small study and the language needs to be more cautious throughout, including in the abstract.

We have repetitively stressed in the abstract as well as the text that this is a small study.

3. There's a brief limitations of the design section, but no section looking at the limitations of the study. This should be added.

The section originally called "limitations of the current design" is now named "limitations". We have also included in this section a full discussion of the study limitations taking into account the reviewers comments.

4. What next? Trials? Fully powered diagnostic tests? Please see our instructions for authors for what we expect from a pilot study.

In the limitations section we have now stated that the next step is a fully powered clinical trial, focusing on diagnostics instead of just event identification

5. Were the cases and controls matched at all?
No. This has now been made explicit in the text.

6. We need a proper explanation of the sample size – what were the numbers based on? We have n=10 subjects and n=20 controls. The article says “The decision on the number of patients was based on obtaining a large enough number of events that would lead to the study goals of 95% confidence intervals for sensitivity and specificity values.”

When the clinical protocol was created to get ethics approval for the study, we were working with an statistician (Ms Pauline Rogers) who was the one estimating the number of subject required. This is what she stated in the ethics application:

"The primary objective is to determine initial estimates of the sensitivity and specificity of the wearable apnoea detection device (WADD) in terms of false-positive and false-negative detections of episodes of apnoea, in comparison with the gold standard current inpatient respiratory and polysomongraphy monitoring. The results will be used to inform the sample size calculations for full evaluation of the devices."

"From clinical experience, it is estimated that the patient group will have a median of 4 episodes of

apnoea per night, range 0 to 6, with 90% (18 of the 20) having at least one episode overnight, and the healthy control group will have a median of 0 episodes of apnoea per night, range 0 to 1, with no more than 5% having one episode of apnoea overnight. The patient group is expected to generate approximately 80 episodes of apnoea. If the apnoea events can be assumed to be independent, this is sufficient to estimate a sensitivity of 90% to within +/- 6.6% (Wald 95% confidence interval for 90%, with a sample size of 80 independent events)."

This has been further clarified in the text. Also, it was noted that Mrs Rogers name had not been added in the acknowledgments, and this has been corrected. She is now retired and this is the only statistical information she gave us.

7. How was the population selected?

Sequential clinical cases, attending for overnight sleep monitoring with question of sleep related apnea. This has been clarified in the text.

The number of patients on which the analysis is done in my opinion is too small to draw definitive conclusions on the diagnostic value field situation.

We agree. The aim of this study was to assess the ability of this new technology to detect apnea events automatically, to compare it an existing used one, and to obtain information that could be used to inform a future fully powered clinical trial. This future clinical trial will focus on diagnosis. All of this has now been fully clarified in the text.

We must insist on comparing the new system with oximetry that remains the main screening tool used yet in routine clinical setting.

Oximetry has been used (both by the clinician to determine whether a reduction in the oronasal, abdominal and chest signals corresponded to hypopnea; and by the Somno system). However, it is not possible to use just oximetry in the context of this study, because oximetry on its own can not detect apnea (i.e. full absence of air flow).

The somnoscreen is recognized for its automatic very low sensitivity analysis and comparison with this software can only show strong limitations of this somnocreen system.

It is true that many people know that the Somnoscreen automatic software does not work properly, but to the authors knowledge nobody has done any attempt to quantify its performance (or any other automatic software for this matter). This is why we believe there is a value on showing these results. Furthermore, we also compare independently with the gold standard.

How to improve the detection of hypopneas? By properly being able to estimate lung volumes from sound. We have already have significant progress in this area. Preliminary results have been already peer reviewed and accepted in EMBC 2014.

I confess also do not believe too much in the future use of this device for the detection of apnea in epileptic patients.
Coming from an epilepsy background we strongly disagree with this. In fact, in the last two months SUDEP Action UK has made speeding up the development of this technology its funding priority for the next year, so that patients can have access to it as soon as possible. Furthermore this has been

directly supported by Samantha Cameron (UK Prime Minister's wife). For more information: https://www.sudep.org/article/samantha-cameron-hosts-sudep

I will add in the "study limitations" that this criterion (tracheal sounds and apnea index rated) does not allow to assess the hypoxic load or autonomic activation and therefore impact the cardiovascular or stroke associated with OSA syndrome.

This has been added.

The correct statistical methods just need to be applied (if applicable) and then described more fully in the manuscript.

Further information regarding the limitations of the study, the number of patients and future work, taking into account the reviewer's comment has been added in the manuscript.

The two methods of assessment are of interest for two different audiences. We understand that a clinical audience would prefer only the method of assessment in which the system is compared to the clinician, but for a more technical (engineering), audience, which is generally more reluctant to accept human quantifications, providing also the other method would also have a significant value. Amongst other things it would show how two very different methods give very similar values. This is one of the reasons why we have chosen to leave both methods.

VERSION 2 – REVIEW

REVIEWER	Frederic Roche, MD PhD Clinical Physiology - VISAS Center CHU Saint Etienne France
REVIEW RETURNED	08-Aug-2014

- The reviewer completed the checklist but made no further comments.

REVIEWER	Matthew Strand National Jewish Health
REVIEW RETURNED	15-Aug-2014

GENERAL COMMENTS	I brought up 2 major issues on the previous review. One response in particular is not sufficient My recommendation is that the authors find a statistician to help them complete this project. It may involve 'minor revision' but is important. Regarding the two methods of assessment, I have no further comments. However, I believe that the longitudinal nature of the data still needs to be addressed in the article. To be clear, I have first included my original comment, then they're response, followed by my new response (counterresponse). **Comment from original review**: Before reviewing the comments (from a previous review) I had also thought that there was a lack of detail regarding the repeated measures. This might impact the performance statistics but will certainly affect the confidence

	intervals. This brings up an even greater issue: how were the performance statistics calculated? If they were based on a longitudinal logistic regression model then the appropriate inference could be performed. Otherwise, the correlation within individuals needs to be accounted for in some fashion, particularly when computing the confidence intervals. The data cannot simply be pooled and treated as independent observations when conducting inference. I'm not sure if that was done or not, but there is no information about it that I could find in the article. Thus, the recommendations here are to (1) properly account for repeated measures in the data when determining performance statistics and related confidence intervals, and (2) describe in the Data Analysis section how this was done. Note that fitting a longitudinal logistic regression model (e.g., employing GEE for a generalized linear model or using a generalized linear mixed model with pseudo-likelihood methods) will give you the ability to handle serial correlation within subjects; such models will also allow you to enter covariates into the model. **Authors' response**: Regarding the issue of pooling the data, we can see the reviewer's point of disagreement with the statistician who advised us to do it this way. We have now however explicitly addressed on the paper that we have obtained these numbers assuming that the events were independent and in some cases they might not be. Having said that, it is worth also noticing that the characteristics of the signal obtained from our sensor changed as much within the same subject (depending on timing, position, external artefacts, etc.) than between different subjects, so the assumption of independence might not be fully accurate only for events that occur in short succession in time. **Reviewer counter-response**: Given that data do involve multiple measures on subjects, I am surprised that there is no mention of it in the Methods. It appears that the authors have assumed that the independence assumption is feasible, and have even used this assumption in order to perform sample size and power calculations. Given that this assumption can be tested more directly by fitting longitudinal logistic regression models, I would suggest that the authors study the assumption further to see how valid it is. If the within-subject correlation is indeed very weak or negligible, then they can proceed as is, and more firmly state that independence assumption is reasonable. Otherwise, they could use the models directly in order to conduct inference. Since the statistician working on the project has since retired, they would probably benefit from recruiting a new one to help. Either way, the repeated measures and how they are dealt with need to be discussed earlier in the

	paper (starting in the Methods), rather than just including it as a short limitation point in the Discussion. I would suggest studying this further, regardless of whether this is considered pilot data or not. Two possible types of correlation that might exist are: serial correlation (e.g., first-order autoregressive), or compound-symmetric (a.k.a. 'exchangeable' using GEE methods). The latter might occur if some subjects tend to generally have more epochs than others. There are different approaches to fit longitudinal logistic regression models; again, I'd recommend a statistician to help describe and choose methods for the researchers.

VERSION 2 – AUTHOR RESPONSE

We have now revised the manuscript further addressing the reviewers point. We have taken three actions:

1- We have explicitly said early on in the methods section that we assumed that the events were independent for the analysis.
2- We have added information regarding the values of individual sensitivities and specificities (i.e. not pooling the data)
3- We have taken 3 random 10 minutes sections of signals for each one of the 30 subjects and run 2700 correlation simulations, to further confirm our observation that the breathing signal changed as much within the same subject as it changed from subject to subject. The maximum correlation coefficient obtained between sections of data from the same subject was 0.05. The maximum correlation coefficient obtained from different subjects was 0.0675.

VERSION 3 - REVIEW

REVIEWER	Matthew Strand National Jewish Health
REVIEW RETURNED	27-Aug-2014

GENERAL COMMENTS	I have read responses from the reviewers and observed the manuscript (number above) and I have no further comments.