PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Classification of accelerometer wear and non-wear events in seconds for monitoring free living physical activity
AUTHORS	Zhou, Shang-Ming; Hill, Rebecca; Morgan, Kelly; Stratton, Gareth; Gravenor, Mike; Bijlsma, Gunnar; Brophy, Sinead

VERSION 1 - REVIEW

REVIEWER	Vincent van Hees Newcastle University, UK
REVIEW RETURNED	06-Feb-2015

GENERAL COMMENTS	The manuscript describes the investigation of a non-wear detection algorithm. General comments: 1. The 3.0 mg threshold for detecting periods of monitor non-wear was copied from a previous study (van Hees 2011 PLoSONE). However, it should be noted that previous study was based on the GENEA accelerometer, while in the current study the authors use the GENEActiv accelerometer. The GENEActiv comes with a different level of noise and therefore requires a different threshold for non-wear detection. The new threshold is likely to be 13 mg, see da Silva et al. in Int J Epidemiology 2014. It should also be noted that the design of the 2011 algorithm was updated in 2013 (see van Hees et al. PLoSONE 2013). As a result the method as evaluated by the authors in the current manuscript is most likely sub-optimal. Please update analysis based on the correct threshold and if possible with the improved design for the non-wear detection. (Note that the code for the existing algorithm is released as open source code in R package GGIR which works with GENEActiv and GENEA data). If you use your own code then this needs to be clarified in the manuscript. 2. There is a fundamental difference between count accelerometry and raw data accelerometry. The authors discuss non-wear detection with count accelerometer quite extensively without highlighting this difference and the potential of better classification with raw data. 3. The authors fail to report whether the accelerometers were worn during sleep, which is important to know in the evaluation of non-

	wear detection. Sleep is the most challenging type of activity for a non-wear detection method. 4. The comparison between accelerometer only and temperature-only method is not standardised. The accelerometer-only method was not developed and evaluated with the same set of equipment, while the temperature-only method were developed and evaluated with the same set of equipment. As a result the evaluation of the accelerometer method will suffer from inconsistencies in equipment while the temperature method does not. 5. The manuscript includes too much spoken language in my opinion. 6. Results are currently limited to typical classification performance indicators. Please expand text with a sentence on number of average misclassified minutes in a day and if possible whether these misclassifications occured during night or daytime. Specific comments: P6L18: "It is reported that..." Please clarify where it is reported. P7L38 – Statement about GENEA and GENEActiv is incorrect. These are different accelerometers (see my comment above). The GENEA was produced by Unilever Discover Ltd, while the GENEActiv was produced by ActivInsight Ltd. The electronics, the colour, the size, the shape, the data format, the dynamic range and the sample frequency are all different. Page 8, 9 and 10 – Description of temperature method lacks clarity; especially the complementary role of the ACF analysis is hard to follow. A much more structured and step-wise description is needed. Page 11 L21 – Numbering of item 4 is unclear because number 1, 2 and 3 where methods, while item 4 is about the evaluation of those methods Page 17 L50 – Novelty claim is unclear. The 2011 and 2013 methods did also rely on time series analysis. The main novelty I see is the use of a temperature data. Please rephrase. Page 18 L27 – The use of a one minute window introduces the risk for incorrect non-wear detection, especially during sleep (see general comment above). Please discuss why this is nevertheless an advantage. Page 18 L50 – These possible limitations for the use of a temperature sensor are concerning. Why do the authors still recommend the usage of temperature sensor? Please clarify.

REVIEWER	Anna Pulakka University of Turku, Finland
REVIEW RETURNED	13-Feb-2015

GENERAL COMMENTS	This article discusses an important issue in accelerometer

	measurement, valid assessment of accelerometer non-wear time. The objective method of assessing non-wear time based on temperature measurement is an important addition to the methodology. Generally, the article is well written and the methods seem valid. However, I have some comments. 1. Methods: 1.1. As the diary report from WT and NWT is the "gold standard" here, it would be important to get some more detailed information on it. I understand it might be difficult to assess validity of the diary, but information on what precision (minutes? hours? was the clock used for diary synchronized with the clock in the accelerometer?) was the WT and NWT recorded would be helpful. Please see also my comment 3.1. on presenting the amount and duration of WT and NWT "bouts". 1.2. Please provide the manufacturer for GeneActiv accelerometer. The size of the unit would also be relevant information as the devices are worn on wrist/ankle and also by children. 1.3. As I am not familiar with the moving average window method, could you please clarify on what increments did the window move, i.e. what was the length of your "event" as in pages 10 and 11 and throughout the text? Did it move at the increments of 60 seconds, 1 second (as in page 13/row 5) or at the sampling frequency, 100 Hz? In case it was 60 seconds, please consider presenting the results as 1 minute units as it is easier to read than seconds (e.g. page 13/rows 5-10, figures S1 and S3). 1.4. It indeed is extremely useful to detect wear time in scenarios when the participant is motionless but still using accelerometer and thus your choice of classifying events with low movement & high temperature as wear-time makes a lot of sense. If I understood correctly, you decided to classify event as NWT in case of low movement & low but increasing temperature (page 10/rows 48-53). As I understand, these events would be rather similar to the first scenario, only that perhaps the participant had just attached the device and that's why the temperature is still low. Could you justify this decision? 1.5. Please provide methods for calculating the classification rate which is presented in Tables 1 and 2. 1.6. ROC curve analysis (page 12/ rows 27-29): does the point nearest to the top left coordinate correspond to the maximum sum of sensitivity + specificity or is either sensitivity or specificity emphasized? Please clarify. 2. Results: 2.1. Generally, it would be useful to see the background information of your participants, such as number, age, sex, length of the wear time, at the beginning of the results. Consider adding a table about this. Some of the text presented in Methods regarding participants could be presented here instead. Especially the age of children/adults and total wear time/participant is missing from the article. Also number of wear/non wear time "bouts" and/or length of wear/non wear periods would help in assessing the validity of

	the diary method. 2.2. Page 13/rows 54-57 and Figure 3: Clearly, the ROC curve indicates good classification accuracy of the chosen T0. For further information to the reader, it might be useful to provide the sensitivity and specificity of the chosen cut-off point as they are on the ROC curve as well as ROC-AUC. 2.3. Page 13/row 21: Figure 3 refers to the ROC curve, the correct reference might be Figure S2. Please clarify. 2.4. In my mind, Table S2 does not bring about much useful information, as it is only an example of one participant and the "whole story" is told in Table 1. In case you decide to keep Table S2, please consider adding all the events, not just events from one participant and percentages to indicate agreement between the method in question and diary report. (Please also mention in the title of the table that "event" refers to seconds.) 2.5. Please provide statistical methods for calculating the p-value presented in page14/row 23 in the Methods. Generally, p-values should be presented with 3 decimals to separate e.g. 0.008 from 0.012 which would both round up to 0.01. 2.6. Table 1: Please use two decimals in Table 1 consistently. 2.7. Does "average" in the titles of Tables 1, 2 and 3 refer to mean or median? 2.8. There seems to be some unnecessary repetition of figures that are presented in Tables 1-3 in the text. Consider reducing the amount of text, one can read the figures from the tables. 2.9. The sentences starting with "No significant difference was found between WT and NWT events…" (page 15/rows 16-19, page 16/rows 27-29, page 17/rows 30-32) are ambiguous. What does the "difference between WT and NWT refer to? Number of WT/NWT events? Classification accuracy of WT and NWT? 3. Discussion: 3.1. Discussion is generally clearly written, although the traditional structure would be to present main results in the first, rather than in the second paragraph. 3.2. The combined temperature and acceleration –based method clearly has benefits. However, I am not totally convinced that these relate to capturing short bouts of PA from children (page 18/rows 5 and 31-33) but the epoch length might be more important in this sense. Even the older, acceleration-based algorithms would capture the short bursts as they would not be zero-counts. Please elaborate. 3.3. You mention that the temperature threshold is UK-specific. How much does staying indoors/outdoors affects the temperature measurement? Does season of the year have an impact? It would be useful to elaborate this in the manuscript, and also mention the season of the year when the study was conducted. 4. Language: language is fluent. However, please check the consistency in the used terminology, e.g. non-wear time/nonwear time/NWT/WNT and tri-axial/triaxial. In addition, the text includes substantial amount of parenthesis. Consider reducing their amount. In some cases the text in parenthesis could be deleted. For

	example, the issues written in page 8/rows 49-57 in parenthesis are repeated, albeit in another words, at the beginning of the following page.

VERSION 1 – AUTHOR RESPONSE

Reviewer Vincent van Hees's Comments

"The 3.0 mg threshold for detecting periods of monitor non-wear was copied from a previous study (van Hees 2011 PLoSONE). However, ... Please update analysis based on the correct threshold and if possible with the improved design for the non-wear detection."
Response: We appreciated the reviewer for providing such information. We have updated the results by using the new threshold [da Silva et al, Int J Epidemiology 2014]. It is found that the overall performance was not improved much by the new threshold, however, the new threshold did improve the specificity and positive prediction value, while the sensitivity and negative prediction value may be degraded to slight extent in some activities.
Revision location: Tables 1, 2 and 3.

" (Note that the code for the existing algorithm is released as open source code in R package GGIR which works with GENEActiv and GENEA data). If you use your own code then this needs to be clarified in the manuscript."
Response: We use our own codes developed in Matlab. It has been clarified in "Data Sharing" statement that the codes are available to readers once upon requested.

"There is a fundamental difference between count accelerometry and raw data accelerometry. The authors discuss non-wear detection with count accelerometer quite extensively without highlighting this difference and the potential of better classification with raw data."
Response: We agree with and appreciate the reviewer for pointing out this important difference. We have added some expressions to highlight it.
Revision location: the paragraph across pages 6 and 7, and the 3rd paragraph on page 19.

"The authors fail to report whether the accelerometers were worn during sleep, which is important to know in the evaluation of non-wear detection. Sleep is the most challenging type of activity for a non-wear detection method."
Response: Good point! Yes indeed, many participants did wear the accelerometers during sleep. But we did not provide detailed quantitative analysis of activities during sleep, because the participants or parents for children were not asked to record the time of sleep periods during data collection. We have reckoned the potential of the proposed method to be applied to analysis of sleeping activity. Some related work has been planned for future research.
Revision location: the 1st paragraph in the Materials and Methods (page 7), and the 2nd paragraph on page 19.

"The comparison between accelerometer only and temperature-only method is not standardised. … As a result the evaluation of the accelerometer method will suffer from inconsistencies in equipment while the temperature method does not."

Response: We are not sure we have fully understood this concern. In this study, we used the same equipments during the evaluation of the different methods. Although the different classification methods work with different mechanisms, the measurements of classification performance are well standardised, for example, the widely used sensitivity and specificity.

“The manuscript includes too much spoken language in my opinion.”
Response: We have tried to elucidate the research problems and methods in easily understood language, as the readers of the BMJ Open are very diversified.

“Results are currently limited to typical classification performance indicators. Please expand text with a sentence on number of average misclassified minutes in a day and if possible whether these misclassifications occured during night or daytime.”
Response: We have added the new results about average misclassified minutes in a day by different methods.
Revision location: the 1st paragraph on page 15.

“P6L18: "It is reported that..." Please clarify where it is reported.”
Response: We have updated our points. Originally it was reported in our unpublished data.
Revision location: the 2nd paragraph on page 6.

“P7L38 – Statement about GENEA and GENEActiv is incorrect…”
Response: We appreciated the reviewer for pointing out such difference. We have revised accordingly.
Revision location: Title of subsection on page 8 and the 1st paragraph on page 8.

“Page 8, 9 and 10 – Description of temperature method lacks clarity; especially the complementary role of the ACF analysis is hard to follow. A much more structured and step-wise description is needed.”
Response: We have tried to clarify the temperature method and used a more structured description.
Revision location: The 2nd paragraph on page 9 to the paragraphs on page 10.

“Page 11 L21 – Numbering of item 4 is unclear because number 1, 2 and 3 where methods, while item 4 is about the evaluation of those methods”
Response: Yes, we agree. We have revised accordingly.
Revision location: The title of subsection on page 12.

“Page 17 L50 – Novelty claim is unclear. The 2011 and 2013 methods did also rely on time series analysis. The main novelty I see is the use of a temperature data. Please rephrase.”
Response: Yes, we agree. We have rephrased accordingly.
Revision location: The 2nd paragraph on page 19.

“Page 18 L27 – The use of a one minute window introduces the risk for incorrect non-wear detection, especially during sleep (see general comment above). Please discuss why this is nevertheless an advantage.”
Response: Yes, on one hand, the use of one-minute window may increase the risk of misclassifying

non-wear events, but on the other hand, the use of one-minute window offers the way of capturing sporadic, short bursts of activity expected from younger participants, in which the whole activity often do not last very long. No doubt, an efficient method is expected to fulfil the tasks of using one-minute window technique while not increasing the risk of incorrect detection of non-wear events. However, majority of the existing methods solely based on acceleration information do suffer from the risk of misclassifying the non-wear events if one-minute window is adopted. Our proposed method combining the acceleration information and temperature information demonstrated impressed performance of correctly detecting non-wear events by using one-minute window.
Revision location: The 3rd paragraph on page 19.

"Page 18 L50 – These possible limitations for the use of a temperature sensor are concerning. Why do the authors still recommend the usage of temperature sensor? Please clarify."
Response: Although we have addressed some limitations of temperature information, clearly the benefits offered by combining temperatures with acceleration information would exceed these limitations, in particular, in the situation of detecting non-wear events in sedentary activity, especially sleeping activity. Importantly, the multi-sensor mobile technologies become more and more popular in new generation accelerometers, there is no reason why one should not use temperature information while it is available already without inducing further costs.
Revision location: The 1st paragraph on page 21.

Reviewer Anna Pulakka's Comments

"This article discusses an important issue in accelerometer measurement, valid assessment of accelerometer non-wear time. The objective method of assessing non-wear time based on temperature measurement is an important addition to the methodology. Generally, the article is well written and the methods seem valid."
Response: We appreciated the reviewer for the supportive comments.

"As the diary report from WT and NWT is the "gold standard" here, it would be important to get some more detailed information on it. I understand it might be difficult to assess validity of the diary, but information on what precision (minutes? hours? was the clock used for diary synchronized with the clock in the accelerometer?) was the WT and NWT recorded would be helpful."
Response: Yes, we have added this information to the manuscript.
Revision location: The 1st paragraph of Materials and Methods on page 7.

"Please provide the manufacturer for GeneActiv accelerometer. The size of the unit would also be relevant information as the devices are worn on wrist/ankle and also by children."
Response: We have added this information to the manuscript.
Revision location: The 1st paragraph on page 8.

"Could you please clarify on what increments did the window move, i.e. what was the length of your "event" as in pages 10 and 11 and throughout the text? Did it move at the increments of 60 seconds, 1 second (as in page 13/row 5) or at the sampling frequency, 100 Hz?"
Response: Yes, the increment of the moving window at each step is 1 second. The size of the moving

window is 1 minute.
Revision location: The last paragraph on page 11.

"It indeed is extremely useful to detect wear time in scenarios when the participant is motionless but still using accelerometer and thus your choice of classifying events with low movement & high temperature as wear-time makes a lot of sense."
Response: We highly appreciated the reviewer for recognising the significance of this research problem.

"If I understood correctly, you decided to classify event as NWT in case of low movement & low but increasing temperature (page 10/rows 48-53). As I understand, these events would be rather similar to the first scenario, only that perhaps the participant had just attached the device and that's why the temperature is still low. Could you justify this decision?"
Response: Yes indeed, the case of low movement & low but increasing temperature does exist when the participant had just attached the device. But in this study, the participants have been asked to record the timings of wear and nonwear after the events of wear/nonwear last for a minimum of 15 minutes. According to our pilot data, this reasonable length of 15 minutes ensured that the temperature could increase to an expected level after the device was worn at least 15 minutes.

"Please provide methods for calculating the classification rate which is presented in Tables 1 and 2."
Response: We have provided them accordingly.
Revision location: The 1st paragraph on page 15.

"ROC curve analysis (page 12/ rows 27-29): does the point nearest to the top left coordinate correspond to the maximum sum of sensitivity + specificity or is either sensitivity or specificity emphasized? Please clarify."
Response: Yes.
Revision location: The 1st paragraph on page 13.

"Generally, it would be useful to see the background information of your participants, such as number, age, sex, length of the wear time, at the beginning of the results. Consider adding a table about this. Some of the text presented in Methods regarding participants could be presented here instead. Especially the age of children/adults and total wear time/participant is missing from the article."
Response: We have summarised and added the background information about the participants accordingly. The total wearing time was added as well. But we didn't record the accurate age of each child and adult, instead we only recognised them as age groups.
Revision location: The last paragraph on the page 13.

"Page 13/rows 54-57 and Figure 3: Clearly, the ROC curve indicates good classification accuracy of the chosen T0. For further information to the reader, it might be useful to provide the sensitivity and specificity of the chosen cut-off point as they are on the ROC curve as well as ROC-AUC."
Response: We have added this information accordingly.
Revision location: The last paragraph on the page 14.

“Page 13/row 21: Figure 3 refers to the ROC curve, the correct reference might be Figure S2. Please clarify.”
Response: We have corrected it accordingly.
Revision location: The 2nd paragraph on the page 14.

“In my mind, Table S2 does not bring about much useful information, as it is only an example of one participant and the “whole story” is told in Table 1.”
Response: We have removed the Table S2.

“Please provide statistical methods for calculating the p-value presented in page14/row 23 in the Methods. Generally, p-values should be presented with 3 decimals to separate e.g. 0.008 from 0.012 which would both round up to 0.01.”
Response: We have revised it accordingly.
Revision location: The 1st paragraph on the page 15.

“Table 1: Please use two decimals in Table 1 consistently.”
Response: We have revised it accordingly.
Revision location: The Table 1 on the page 15.

“Does “average” in the titles of Tables 1, 2 and 3 refer to mean or median?”
Response: Mean.
Revision location: The title of Table 1 on the page 15.

“There seems to be some unnecessary repetition of figures that are presented in Tables 1-3 in the text. Consider reducing the amount of text, one can read the figures from the tables.”
Response: We have removed some repetitions.
Revision location: The Table 1, Table 2 and Table 3.

“The sentences starting with “No significant difference was found between WT and NWT events…” (page 15/rows 16-19, page 16/rows 27-29, page 17/rows 30-32) are ambiguous. What does the “difference between WT and NWT refer to? Number of WT/NWT events? Classification accuracy of WT and NWT?”
Response: We have clarified them accordingly .
Revision location: The 1st paragraphs on pages 15, 16, and 17 respectively.

“Discussion is generally clearly written, although the traditional structure would be to present main results in the first, rather than in the second paragraph.”
Response: We have further revised it .
Revision location: The paragraph of Discussion across pages 18, 19.

“The combined temperature and acceleration –based method clearly has benefits. However, I am not totally convinced that these relate to capturing short bouts of PA from children (page 18/rows 5 and 31-33) but the epoch length might be more important in this sense. Even the older, acceleration-based algorithms would capture the short bursts as they would not be zero-counts. Please elaborate.”

Response: We appreciated the reviewers for recognising the benefits of combining temperature and acceleration. We do acknowledge that the acceleration-based algorithms could capture the short bursts once they would not be zero-counts. But what we meant by the advantage of the proposed method using 1-minute moving window is that the proposed method aims to capture the sporadic, short bursts of activity expected from younger participants, in which such activity often do not last very long. Because the acceleration-based algorithms often rely on very long block of data (in the order of 30 minutes, many in 1 or 2 hours), so it would be difficult for them to classify the sporadic, short bursts of activity that does not last long enough, which often happen in children's activities.
Revision location: The 3rd paragraph on page 19.

"You mention that the temperature threshold is UK-specific. How much does staying indoors/outdoors affects the temperature measurement? Does season of the year have an impact? It would be useful to elaborate this in the manuscript, and also mention the season of the year when the study was conducted."
Response: Our study didn't consider the variations of temperature in different situations. We have pointed the potential impact of these variations.
Revision location: The paragraph across on pages 20 and 21.

"Language: language is fluent. However, please check the consistency in the used terminology, …"
Response: We have checked the consistency across the whole manuscript. But we have to keep the texts in parenthesis on page 9, as they elucidate the different internal associations between temperature observations.

VERSION 2 – REVIEW

REVIEWER	Vincent van Hees Newcastle University
REVIEW RETURNED	24-Mar-2015

GENERAL COMMENTS	1. Quote: "We are not sure we have fully understood this concern. In this study, we used the same equipments during the evaluation of the different methods. Although the different classification methods work with different mechanisms, the measurements of classification performance are well standardised, for example, the widely used sensitivity and specificity." Response: What I meant was that your accelerometer only method relies on an externally provided threshold while you train your CTA method with your own devices. This could in theory explain the different findings. In order to address this issue I recommend that the authors verify that 14mg actually makes senses for their set of accelerometers. Please collect data under non-movement conditions (accelerometers on table) and check the standard deviation in the signals and use this to verify whether the 14 mg threshold is appropriate for your accelerometer equipment. Also report that you did this test in the text. 2. Quote: "Yes, on one hand, the use of one-minute window may

	increase the risk of misclassifying non-wear events, but on the other hand, the use of one-minute window offers the way of capturing sporadic, short bursts of activity expected from younger participants, in which the whole activity often do not last very long. No doubt, an efficient method is expected to fulfil the tasks of using one-minute window technique while not increasing the risk of incorrect detection of non-wear events. However, majority of the existing methods solely based on acceleration information do suffer from the risk of misclassifying the non-wear events if one-minute window is adopted. Our proposed method combining the acceleration information and temperature information demonstrated impressed performance of correctly detecting non-wear events by using one-minute window. Revision location: The 3rd paragraph on page 19." Response 1: The authors are confusing the duration of activities with the duration of non-wear periods. Why would a child or an adolescent take the monitor off for shorter periods of time compared with an older person? It is most likely that neither young nor old individuals will ever take the monitor off for periods shorter than 15 minutes. Therefore, I personally think that a 5 or 10 minute window would be more convincing. You may want to reconsider this. Response 2: Please clarify what window size was use for the accelerometer only technique? Do I understand it correctly that the authors also used a one minute window for this? If yes, then this means that the resemblance with my own 2011 algorithm and its 2013 enhancement is very limited. Please make sure that this is clarified in the text. 3. "Although we have addressed some limitations of temperature information, clearly the benefits offered by combining temperatures with acceleration information would exceed these limitations, in particular, in the situation of detecting non-wear events in sedentary activity, especially sleeping activity. Importantly, the multi-sensor mobile technologies become more and more popular in new generation accelerometers, there is no reason why one should not use temperature information while it is available already without inducing further costs. Revision location: The 1st paragraph on page 21." Response 1: Low cost and availability of information are in my opinion no strong justifications for using information, especially when there is a severe risk for introducing measurement error. A much strong justification is increased accuracy that can be gained. Further, with the use of more information sources there is an increased risk for overfitting a model to a specific experimental context. Response 2: The authors conclude that accerometer and

	temperature combined perform relatively better than an accelerometer only method. The presented findings are specific to the two methods as implemented. A modification to the accelerometer only method, may well result in entirely different performance. Please ensure that conclusion statement emphasizes that these findings are limited to your current implementation of the algorithms and may not generalize to alternative non-wear detection algorithms. In the physical activity research scientists sometimes lack the skill to understand that different algorithms can produce different outputs. They usually naively assume that it is all about good and bad sensors types. Therefore, it is particularly important that it is emphasized that these findings are specific to the algorithms as implemented. Response 3: The method comparison is currently limited to a dataset with 27% of non-wear time mostly fractured over the day as a result of the study design. Therefore, one could argue that a different non-wear detection method for accelerometer only could perform fine if non-wear time is assumed to be predominantly non-fractured and last for substantial periods of time. This needs to be acknowledged in the discussion section.

REVIEWER	Anna Pulakka University of Turku, Finland
REVIEW RETURNED	12-Mar-2015

GENERAL COMMENTS	I appreciate the authors for addressing the issues raised up in the review. I think the authors have appropriately clarified most of the issues and I only have two minor comments: 1. In the authors' response to comment about the low movement & low but increasing temperature, they mention that "…the participants have been asked to record the timings of wear and nonwear after the events of wear/nonwear last for a minimum of 15 minutes..." This raises a concern that the first 15 minutes of each wear or non-wear event was not included in assessing the performance of the algorithm. If this is the case, it should be clearly explained in the methods as it has implications for interpreting the results. This also somewhat contradicts the current statement on page 7: "Participants… kept diaries of the exact timings of accelerometer wear and removal…" 2. Methods for calculating the classification rate and comparing performance characteristics, which are now located on page 15 in results, would be better situated in Methods-section, e.g. at the last paragraph where other statistical methods are explained.

VERSION 2 – AUTHOR RESPONSE

Reviewer Anna Pulakka's Comments

"I appreciate the authors for addressing the issues raised up in the review. I think the authors have appropriately clarified most of the issues"

Response: We appreciated the reviewer for the supportive comments.

"In the authors' response to comment about the low movement & low but increasing temperature, they mention that "…the participants have been asked to record the timings of wear and nonwear after the events of wear/nonwear last for a minimum of 15 minutes..." This raises a concern that the first 15 minutes of each wear or non-wear event was not included in assessing the performance of the algorithm. If this is the case, it should be clearly explained in the methods as it has implications for interpreting the results. This also somewhat contradicts the current statement on page 7: "Participants… kept diaries of the exact timings of accelerometer wear and removal…"."

Response: We understand this concern. We have added one sentence to clarify it.
Location: The 1st paragraph on page 12.

"Methods for calculating the classification rate and comparing performance characteristics, which are now located on page 15 in results, would be better situated in Methods-section, e.g. at the last paragraph where other statistical methods are explained.."

Response: We have revised accordingly.
Location: The paragraph next to the last on page 12.

Reviewer Vincent van Hees's Comments

"What I meant was that your accelerometer only method relies on an externally provided threshold while you train your CTA method with your own devices. This could in theory explain the different findings. In order to address this issue I recommend that the authors verify that 14mg actually makes senses for their set of accelerometers. Please collect data under non-movement conditions (accelerometers on table) and check the standard deviation in the signals and use this to verify whether the 14 mg threshold is appropriate for your accelerometer equipment. Also report that you did this test in the text."

Response: We appreciate the reviewer's recommendation, but we would point out that although our study used the externally provided threshold, this threshold was actually provided in previous study [27] with the same devices (GeneActiv) as ours. Also we have tested the 14 mg, but the overall performance was not improved by it, even the sensitivity is slightly worse. So we still use the current threshold as suggested in your previous review report.

"The authors are confusing the duration of activities with the duration of non-wear periods. Why would a child or an adolescent take the monitor off for shorter periods of time compared with an older person? It is most likely that neither young nor old individuals will ever take the monitor off for periods shorter than 15 minutes. Therefore, I personally think that a 5 or 10 minute window would

be more convincing. You may want to reconsider this."

Response: We did misunderstand this point. We have clarify that the purpose of setting at least 15 minutes is to make sure the accelerometer temperature measurements correctly reflect the changes of events – wear and non-wear respectively. We agree that neither young nor old individuals will ever take the monitor off for periods shorter than 15 minutes, which is just the reason why we set such a lasting period (15 minutes). This makes sure that the proposed method is applicable to activities lasting longer than 15 minutes. We would not like to choose 5 or 10 minutes, as we are worried such shorter period of time may not be enough to observe the temperature changes of wear and non-wear events.
Location: The paragraph next to the last one on page 8.

"Please clarify what window size was use for the accelerometer only technique? Do I understand it correctly that the authors also used a one minute window for this? If yes, then this means that the resemblance with my own 2011 algorithm and its 2013 enhancement is very limited. Please make sure that this is clarified in the text."

Response: Yes, we use a one minute window for the accelerometer only algorithm, which is different from the ones in previous studies mentioned. We have clarified this.
Location: The 1st paragraph on page 12.

"Low cost and availability of information are in my opinion no strong justifications for using information, especially when there is a severe risk for introducing measurement error. A much strong justification is increased accuracy that can be gained. Further, with the use of more information sources there is an increased risk for overfitting a model to a specific experimental context."

Response: Good point! We have revised accordingly.
Location: The end of last paragraph in Discussion section (page 21).

"The authors conclude that accelerometer and temperature combined perform relatively better than an accelerometer only method. The presented findings are specific to the two methods as implemented. A modification to the accelerometer only method, may well result in entirely different performance. Please ensure that conclusion statement emphasizes that these findings are limited to your current implementation of the algorithms and may not generalize to alternative non-wear detection algorithms. In the physical activity research scientists sometimes lack the skill to understand that different algorithms can produce different outputs. They usually naively assume that it is all about good and bad sensors types. Therefore, it is particularly important that it is emphasized that these findings are specific to the algorithms as implemented."

Response: We appreciate the reviewer for pointing out this scenario. Although our study was based on the methods of temperature only, acceleration only and combining two data together (CTA), there is great potential is to combine this sample temperature only method to other acceleration only methods for accelerometers with temperature sensors. We have added explanations about this point.

Location: The 1st paragraph on page 20, and Conclusion section.

“The method comparison is currently limited to a dataset with 27% of non-wear time mostly fractured over the day as a result of the study design. Therefore, one could argue that a different non-wear detection method for accelerometer only could perform fine if non-wear time is assumed to be predominantly non-fractured and last for substantial periods of time. This needs to be acknowledged in the discussion section.”

Response: We appreciate this argument, but we don’t think predominant non-wear time would change the nature of the addressed problem: combining acceleration and temperature information has much to gain. Importantly, predominant non-wear time is less uncommon than the predominant wear time. Our study addresses the common scenario of wear and non-wear events with predominant wear time.