Author's response to reviews

Title: Estimates of adherence and error analysis of physical activity data collected via accelerometry in a large study of free-living adults

Authors:

David R Paul (dpaul@uidaho.edu)
Matthew Kramer (Matthew.Kramer@ars.usda.gov)
Kim S Stote (Kim.Stote@ars.usda.gov)
Karen E Spears (kspears@cabnr.unr.edu)
Alanna J Moshfegh (Alanna.Moshfegh@ars.usda.gov)
David J Baer (David.Baer@ars.usda.gov)
William V Rumpler (Bill.Rumpler@ars.usda.gov)

Version: 3 Date: 12 December 2007

Author's response to reviews: see over
December 10, 2007

Dear Editor of BMC Medical Research Methodology,

On behalf of my co-authors, I would like to take the opportunity to address suggestions to our manuscript “Estimates of adherence and error analysis of physical activity data collected via accelerometry in a large study of free-living adults” (MS# 4803636301525387) made by the reviewers. We appreciate the time it took for them to provide such thorough and helpful suggestions.

Please let me know if you have any questions.

Thank You,
Dave Paul

Reviewer#1:

General

1) This is a well-written, concise, and interesting study that addressing a well-known problem that is relevant to many users of accelerometer data and comes up with some practical solutions. It is rather unique, although its practicality is limited by the fact the authors have taken a rather pessimistic approach of simulating the removal of a total of 10 hours of data (in 10 groups of 1 hour).

> This reviewer has slightly misinterpreted the removal simulations. The 10 hours of removals refer to removing 10 hours of data from the active part of the day, not including sleep. In addition to the 10 groups of 1 hr, we removed a 2 hr, 3 hr, to up 10 hr time blocks to examine imputation effects.

-------------------------------------------------------------------------------

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)

There do not appear to be any major problems that must be revised.

-------------------------------------------------------------------------------

Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)

Page 5 Line 88 – could replace “intervals of time” with “epochs”

> Correction made

Page 5 Line 97 – delete the word “long”. It is true removing the AM will produce a string of zeros, but whether it is long or short depends on how long it is removed. Simply removing the AM will not automatically produce a long string of zeros (a somewhat pedantic suggestion I admit).

> Correction made
Page 7 Line 121-2. That the computer program predicted waking and sleeping only to within 12 and 41 minutes of the reported times seems very large (especially the 41 sleeping time), yet no comment is made. How did these time compare with manually/visually estimated times from the dataset (which is fairly easy to do after a bit of experience and stated in Line 261-2). I would be very hesitant accepting such a large computer-generated sleep time error compared to the subjective report unless it could be shown to be much smaller using manually/visually?

> Following this reviewers suggestion, we went back to data and estimated the wake and sleep times manually/visually in a blinded manner. The analysis indicated that the wake/sleep times (between the computer program and manual/visual approach) were different by an average of 6.0 min (wake times) and 34.6 min (sleep times). This discrepancy with the reported sleep times (41 min) is comparable. Despite these differences, the average PA with the manual/visual method resulted in 230.1 counts/min (compared to 228.2 with the computer program), with an average between-day coefficient of variation between the two approaches of 0.7%. Given these small differences, it is very unlikely that the conclusions of the study would be significantly influenced by the use of our computer program to predict sleep. The advantage to an automated program is that the wake and sleep times are estimated much more rapidly and remove any subjective judgments on the part of an investigator. More information about this concept has been included in the discussion.

Discretionary Revisions (which the author can choose to ignore)

Page 2 Line 30 – the authors state no study on adults proposes a method to minimize the effects of accelerometer removal (presumably as a justification of their research). Technically this may well be true, but the Catellier et al (MSSE) paper does this with data from children and I am not aware of any reason to expect why the methodology for imputation of children’s accelerometry data would be any significantly different to that of an adult. Hence the authors might wish to give Catellier et al a bit more credit?

> There are a number of possible goals of imputation, one reason there are so many imputation techniques. Our goal was to produce a “best” estimate in the minimum mean squared error sense. This does not appear to have been Catellier et al.’s goal, as they wanted to preserve the original variances of each student’s data (as well as impute a reasonable estimate for the missing data), since they sample from a distribution, rather than use the point estimate of its mean. However, they do not detail how they derive the parameters of the distribution, and any resulting characteristics of it. Our methodology is clearly simpler and easier to implement, as its main characteristic is that it does not borrow information across subjects. This conforms to our belief that information from other subjects (or even other days of the same subject) do not necessarily improve estimates of the current subject for the current day, but that nearby data values of the current subject are informative. As far as we know, while both Catellier et al. and our study demonstrate that imputing missing values improves daily estimates, our methods for imputing the values are unrelated. Unfortunately, the Catellier et al. technique cannot be reproduced based on what is published in their paper (critical pieces of information to reproduce the
procedure weren’t included), and, despite several requests, she did not supply the necessary information to us.

Page 6 Line 100-101 – it is not clear why the authors defined 20 min of zero to represent a non-wear event, but then go on to simulate a full 60 minutes of data loss? I am not sure where the 20 non-wear definition aids the study?

> Overall, the use of the 20 min criterion is to serve as an estimate of adherence that does not rely on subject self-report. The difficulty with using the 20-min criterion for removals is the notion that not every investigator agrees with this “cut-off”. In response to a similar question by another reviewer, we estimated adherence using a number of different combinations of time; the cut-off used makes little difference (see “Patterns of AM Adherence In A Large Dataset of Adults”). The 60 min criteria was chosen to avoid as much ”controversy” as possible.

Page 9 Line 183 – is a reference needed for the AIC criterion?

> We now cite the book by Burnham and Anderson

Page 10 Line 202 – I think 10 hours is becoming the common default value

> We have not come across any references to indicate the strength of any particular time cut-off.

Page 11 Line 223-5 – not clear why this result occurred. Does this need a slightly longer explanation?

> We did not want to dwell on this further, because it is an “artifactual” effect.

Figures: consider increasing the font sizes for axes and legends – maybe hard to read for some people.

> Correction made

Reviewer#2:
General
The manuscript has lots of potential, but it is not focused. I would suggest that the authors focus on the utility of their findings for future researchers. Here are some comments:
1) In general (most cases) a researcher is interested in grouping his study subjects. Hence the ranking of physical activity levels is the crucial whereas the exact amount of activity is not the aim. Hence the authors need to show what are the implications of excluding the hours without the device on the ranking of the physical activity distribution in a study.
This is an excellent point. However, this paper is already rather complex for its intended audience, and adding additional analyses will only increase the complexity (and length). This is something we could and would like to address in a separate publication. This brings up the question of why one would be satisfied with rankings when good estimates of PA are available---surely they would be more useful as covariates or as a grouping variable than ranks.

2) The subjects are a select group of people who choose to participate. What is the impact on the findings? How could this be applicable in the general population?

This is another important consideration that is rarely discussed in the literature. The highly adherent study subjects had an average body mass index of 26.7 kg/m\(^2\) (± 4.7), with 65.7% of them having BMI’s greater than 25. This range compares well with the NHANES database average of 66.3% (Ogden et al. 2006). The average physical activity level (total energy expenditure/ resting energy expenditure) was 1.66 (±0.29), which is slightly lower the Institute of Medicine (2002) database, but comparable. The between-subject coefficient of variation (CV) for the accelerometer data was 30.3%, demonstrating that there was a wide range of physical activity between the subjects. All of these important measures indicate that the population represented in the simulations were similar to the general population. We have also included more information about the demographics of the subjects.

3) To focus on "removing the data observed during sleep" is not a major issue. The focus should be on removing the data during the day (not sleep). There is a big difference and the authors could use their self-reported data to provide useful information on that impact.

We respectfully do not agree with this notion, and our data demonstrate why we feel this way (Table 3). If every study subject in a particular study woke up and went to sleep at the same time (day in an day out), sleep data would likely be insignificant. However, this is clearly not the case in real life. Differences due to sleep alone produce a CV of 3.1%. The second paragraph of the Discussion describes these issues in greater detail.

4) How could somebody wear this for 24 hours? If true these are outliers and need to be excluded (very unlikely in real life). The same should apply for folks who used the device for a very short time.

It is true that not many of our 500+ subjects wore the monitors for 24 hrs/day. However, these subjects are not “outliers” per se, and correcting for the sleeping hours (as proposed) will actually normalize those differences.

---

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)

Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)
1) In line 82, what metabolic diseases? Please specify.

>Correction made

2) Line 94, reference 27 is from 1983.

> This reference is correct as it is listed in the original manuscript.

3) Lines 97 to 101. What if I was wearing the device for 20 minute but sitting on my desk (for the last 30 minutes, my score would be zero as I am doing this review)? The authors could refine their method by a combining the use of the zeros and the self reported data.

> We acknowledge the weakness of using a particular cut-off for inactivity. However, the monitors are quite sensitive and it is difficult to sit still for prolonged periods of time without triggering some activity. Although this is an imperfect approach, we are not confident that using self-reported data would increase the precision of the estimates.

4) Lines 108-109. How could you judge your criteria by using your outliers? Need a better justification or strategy.

> We are a little uncertain of the question, but we will try to answer as best we can. The adherence estimates (approximately 15 hours/day for 12 days) were estimated from the 500+ subjects (not just the highly adherent ones). The purpose of selecting the highly adherent subjects (24 HR) for the simulations was that they represented a “template” of daily physical activity we could use to simulate what would have happened if they have removed their monitors. You cannot use the entire database because we have no idea about the lost data when the less adherent subjects removed their monitors.

5) Line 117. What "preliminary analyses"?

> Early on in the data analysis phase of this study, we wished to use the self-reported records. When checking first to see if they subjects accurately recorded their wake/sleep times, we realized they did not. Also, subjects occasionally forgot to write the sleep times altogether.

Reviewer#3
General
This paper explored estimates of adherence and error analyses of physical activity data collected by accelerometry in free-living adults. The authors describe the adherence in their study population and examine methods to assess the inherent error associated with nonwear and how imputation procedures can be used to reduce this error. The paper is certainly of value for those using accelerometers to assess free living physical activity levels in adults. The paper raises some important issues, though most are not unique. The paper relies heavily on the Catellier et al. (2005) paper, and for the most part appears to use an adult population to examine the same general issues. These issues were also addressed by Esliger et al. (Journal of Physical Activity
and Health 3:366-383, 2005). The paper provides some addition to the literature in this area however but with some additional work it’s contribution could be significantly enhanced.

> We appreciate the thoughtful comments from this Reviewer. During the process of our data analyses, we have struggled with many of the questions raised, and are happy to see that other scientists are dealing with the same issues. The Esliger paper is not listed on PubMed, so we were unaware of its existence.

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)

1. line 66-67 - the National Health and Nutrition Examination Survey in the U.S. have adherence rates available through their publicly available dataset on several thousand Americans, representative of the general population in the U.S. Insufficient information is provided by the authors to get a sense of the representativeness of the population used and the likelihood of the findings being generalizable to other adult populations with different ethics mixes, ages, climate situations, social norms, etc. In this regard, some comparison to a representative dataset, or at least some discussion on this issue would help inform readers of the potential utility of the recommendations provided in the manuscript. Please provide some information on demographics of the sample (socioeconomic status/mix; country of data collection).

> A table describing the demographics of the subjects has been included.

2. line 99-101 - 20 consecutive zeros may or may not be appropriate – how sensitive were the findings to this assumption? What if this rule was 30 consecutive zeros? 40? If the results are very sensitive to this, the more important factor becomes this decision rule, not the type or use of imputation. Some assessment of this would be helpful.

> This is an issue that is rarely discussed in the literature, and hopefully some future studies will address this issue more specifically. The adherence (hours/day ± SD) for the 15, 20 (what we used), 30, and 40 min criteria are 15.5 (3.3), 15.8 (3.4), 16.3 (3.6), and 16.8 (3.8). Therefore, the criterion of choice has little effect on the estimates of adherence. We have added some additional information to reflect this in the manuscript.

3. Similar to the point above, a program was created to predict wake and sleep time because the authors do not trust the diaries. These values differ significantly from the sleep time recorded by the participants (41 min). How sensitive are the subsequent procedures to this program? What if it were modified to more closely approach what respondents recorded? What if you accepted the recorded times of the respondents? Would these changes strengthen or weaken the effects of imputation?

> In response to this question by Reviewers 1 and 3, we went back and estimated the wake and sleep times manually (by visually inspecting each daily dataset) in a blinded manner. We found the differences between the computer program and manual approach to be 6 min (wake times) and 34.6 min (sleep times). This discrepancy in sleep estimates is comparable with the reported sleep times (41 min). Despite these differences, the average PA with the manual method was 230.1 counts/min (compared to 228.2 with the computer
program), with an average daily coefficient of variation of 0.7%. Given these small differences, it is very unlikely that the conclusions of the study would be significantly influenced by the use of our computer program to predict sleep. The advantage to an automated program is that the wake and sleep times are estimated much more rapidly and remove any subjective judgments on the part of an investigator. We have included some more information to reflect this in the manuscript.

4. Are the results similar between men and women? Between weekdays and weekend days? Again, trying to get at the generalizability or universality of the recommendations made by the authors. Greater efforts to test how robust their recommendations are should be made.

The estimates of adherence by gender were virtually identical (15.81 vs. 15.85 hrs/day). Least squares means for adherence by day-of-week (Sunday to Saturday) were 14.9, 16.1, 16.1, 16.2, 16.3, and 15.0 hrs/day. Adherence was lower for Sunday vs. weekdays, and Saturday vs. weekdays. Sunday and Saturday were not different from each other, and the weekdays were not different from each other. These trends have been included in the manuscript.

5. The choice to use an exponential decay model is fine, but needs to be tested against other models (when you assess only one, how can you recommend it? It was only compared to doing nothing). Also, physical activity behaviour is only somewhat autocorrelated with neighboring times (for example most work outs start and end rather abruptly). The first figure in figure 2 shows the noise associated with most movement behaviours. Other models should be attempted to justify the choice made.

Since we only use the subject’s data from that day for imputation, about the best that can be done is estimate the one autocorrelation parameter. Our experience in time series modeling suggests that estimates from other one parameter models based on < 24 observations would be statistically indistinguishable (even for models as different as MA(1) and AR(1)). There are certainly a large number of potential models---they would differ by how much relative weight is given to neighboring observations. However, for pretty much any reasonable model chosen, more weight would be given to observations closest to the missing data, so the most important question is how fast the autocorrelation dies off---this is the parameter that we estimate using the kriging (AR(1)) method. The imputed values for the model where the autocorrelation dies off exponentially with increased distance would be almost identical to those following a spherical or Gaussian model, given that the model parameter for how fast the autocorrelation dies off is estimated from the same data set. One would probably need > 100 observations to distinguish between models (we have < 24). More complex models (i.e. with more parameters) could not be supported by one day’s worth of data.


The typical pattern of daily physical activity takes on a “bell-shape”, like the first figure in Figure 2. However, due to the lifestyles of the subjects (shift workers, for instance), the patterns of PA can look very different; bi-modal shapes (two bell-curves in
the same day), as well as positively and negatively skewed curves, can occur. These patterns have no effect on any other measures in this study, however. The only place where these issues can be problematic is during measurements of sleep; fortunately the program we developed takes these different wake/sleep cycles into account.

7. Many if not most studies using accelerometers determine the number of minutes of moderate and vigorous PA as their primary outcome variable. What happens to these results when the recommendations are applied? This would be important to assess and would likely be affected by some of the recommendations and the reduced variation resulting from the recommended data imputation procedures.

> The procedures have no effect on estimates of moderate/vigorous PA, since in order to measure these variables, you sum the minute by minute data. The imputations are performed on hourly means, so the procedures influence only the daily means.

8. How were saturation values dealt with (counts of 32767)? These spurious values can have a significant effect on average count values (obviously).

> Another excellent point rarely discussed in the literature; the manufacturer used to deny that these values occurred (not sure if they still do!). Similar to what was discussed in the Esliger 2005 paper, we scanned the database for these values. One of the datasets was composed entirely of “32767”; we forgot to mention the loss of the data, and have revised than manuscript to reflect this. Other than the one subject, there were not very many of them; when they did occur, we consulted the activity logs and looked at the neighboring data points, then imputed a value that appeared to be representative of what the point should have been.

9. In the absence of comparsion analyses the paper cannot justify recommending the kriging (or any other) procedure (line 258). This whole paragraph is based on weak support.

> We disagree (see above). We appreciate that there may be other approaches. The section discussed here states “We suggest the following approach to minimize the effects of AM removals (Simulation D)“. Nowhere in the manuscript does it say that this approach is “the best”; we merely describe an approach and quantify the improvements in the estimates of PA.

Note that, even if we knew the “true” model (that is, both the true form of the model and the true parameter values), the imputed values using them would still not exactly replace the missing data (and might be just as far from them as estimates using the kriging technique) because there is so much noise in activity monitor measurements due to the nature of the device, the sometimes poor translation of PA to monitor counts, and the stochastic nature of hourly PA itself. Our main point here is that any reasonable imputed value is better than discarding or accepting zeroes for the missing data. We don’t believe that any method (that we know of) will produce perfect estimates of the missing data, and, in fact, that the different methods will have roughly the same amount of error (this is a theoretical argument that we do not give in the paper, based on the data having a large $\sigma^2$, and can be found in many books on time series analysis under sections describing one-step
ahead forecast accuracy). As pointed out by Catellier et al., what are really needed are good covariates to improve the estimates, which were not available to us (nor, apparently, to them). These would be variables such as heart rate, energy expenditure, detailed descriptions of activity during the missing times, etc. Of course, if these were available, they would likely better predict PA than using activity monitors.

Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)

1. Please provide research hypotheses.

>Correction made.

2. What time of year was the data collection - discuss how this may affect findings.

>Correction made.

3. Were incentives provided to the participants (e.g. honorarium for wearing the AM)? Discuss how this may impact results and applicability of findings.

>The subjects received an honorarium, but not for wearing the AMs, per se. The PA monitoring was part of a larger study, so the honorarium was tied to the subjects reporting of food intake and other measurements. This information has been included.

4. Please describe how the accelerometers were calibrated before use in the study.

>Correction made.

5. line 120 - citing reference 28 this way reads like these authors developed SAS, which I don't think they did - they can probably clarify by saying "developed a computer program using SAS (spell out) ..."

>Correction made.

6. Simulation D should be added as another figure in figure 2.

>The purpose of Figure 2 was to give the reader an idea of what the minute-by-minute data looked like with the simulations. Simulation D is based on hourly blocks on time, not minute by minute data.

7. line 178: It is stated that TEE was calculated using the equations by Weber et al. (2001). It would be more appropriate to reference the original papers (cited by Weber et al. in their methods) which describe both the equation used to calculate energy expenditure from CO2 production (Weir, 1949) and the sampling protocol used: multi-point method (Schoeller 1983,
1988). Also, the manufacturer details of the isotope ratio mass spectrometer used should be provided in brackets (page 9, line 177).

> The Weber et al. calculation is a “streamlined” equation based on commonly used calculations in the literature. However, we have included additional details to clarify any confusion.

8. Please provide a reference for Akaike's Information Criterion.

> We now cite Burnham and Anderson.

9. line 266 - define "profound understanding of imputational statistics"

> Imputational statistics is a vast field and can be very theoretical. Many imputation techniques do require a good theoretical understanding of the method, in part to aid in diagnosing when things go wrong. Kriging, as we have implemented it, does not require this sort of knowledge. In the worst case scenario, when the estimation fails, the imputed value would be the mean (as it would be if the autocorrelation parameter is estimated to be zero). Many (most) imputation techniques also require a fair amount of programming skills. Kriging (or using an AR(1) model for imputation) is widely implemented in commonly used statistical software packages, so might only require an extra line of code or a checkbox ticked off to get the desired estimates.

10. lines 280-281 - where are the results to support this?

> We presented this at the FASEB conference a number of years back, but forgot to place the reference here. We have made this correction.

11. line 435 - adherence of what?

> Correction made

12. line 440 - add "averages of each individual" before "hours"

> This revision would not be entirely correct. This is a frequency distribution of the daily adherence averages for the entire database.

13. Figure 1 - are these averages per person over the days worn by each person - or are these based on each day of each person? If it is not the latter, it should be. It should be by person/day - this needs to be made clear in the paper.

> Correction made

14. Table 1 - I do not understand the "# of simulations/ subject" column – what does this mean/refer to?
This table demonstrates how many simulations were conducted on each subject. For simulations A1-C1, seven simulations were carried out (one for each of the seven days). The remaining simulations, there were 19 simulations per day (10 x 1 hr, 1 x 2 hr, 1 x 3 hr, etc…) times seven days (133 simulations).

Discretionary Revisions (which the author can choose to ignore)
1. line 88 change "read" to "store"

>Correction made.

2. line 89 put "(epoch)" after "time"

>Correction made.

3. line 92 13-15 continuous(?) days?

>Correction made.

4. Please define the REE methods a little more. What metabolic cart was used?

>We made our own metabolic measurement systems, which requires a more complicated explanation than we think is necessary for this paper.

Minor points: Any control for physical activity in previous 24 hours? Any control for smoking or alcohol intake?

>No strenuous activity the day before the REE measurement, and no controls for smoking or alcohol intake.

5. line 205 include "after" before "omitting"

>Correction made.