Author’s response to reviews

Title: Assessing the precision of a time sampling based study among GPs: balancing sample size and measurement frequency

Authors:

Daniël Van Hassel (danielvanhassel@hotmail.com)

Lud van der Velden (l.vandervelden@nivel.nl)

Dinny de Bakker (d.debakker@nivel.nl)

Lucas van der Hoek (l.vanderhoek@nivel.nl)

Ronald Batenburg (r.batenburg@nivel.nl)

Version: 1 Date: 30 Sep 2017

Author’s response to reviews:

Reviewer reports:

Reviewer #1: Dear Authors,

I enjoyed reading your paper very much and the novality it brings to the topic of time sampling.

There are a few issues to elaborate a bit further:

1. Page 9 - the measurement evaluation line 20 - you have explained well the simulation and fixed it to 40 hours per week and the real life situation is as explained on Table 1 is about 42.9 which can also go up to 51.9hrs for self-employed male GPs - do you think your estimated SE measurements can be sensitive to the set value of 40 hours?

Response:

The 40- hours working week is fictional and the simulations gain insight in measurement fluctuation. We further explained this in the methods section and in the limitations of the discussion section.

We conducted the simulations on the 40- hours working schedule, as earlier research showed that a large group of GPs is working approximately during these moments of a week. The methods section (analyses, step 2) includes a reference to this publication in which the results are reported. We believe that this average working schedule will provide us the most proper insight
into the amount of measurement fluctuation. Nonetheless, we conducted additional simulations on other types of working weeks as well. In most cases this resulted in small differences. The results are now reported in the methods section (analyses, step 2), in the limitations of the discussion section and in the appendix we added a table showing the results of the simulations.

2. Page 11 - the total 95%-CI +/- equation - here it is a combined estimate of the SE sample + SE measurement - can the author provide an explanatory note just to reassure the reader on the relative contribution of the two types of error - the way the authors treated the measurements is not as random as usually the case with repeated measurements which do not usually have an ordinal difference between them - for example the results say that for a sample of 50, the total 95%-CI +/- can reduce from 3 hours to 1 if we take measurements every 30 minutes (in an 8-hour working day) compared to every 3 hours. Furthermore according to Figure 2 (bottom figure), the measurement fluctuation effect (i.e. using every 30 minutes vs every 3 hours) can only optimize itself to about 30 minutes when we have samples of 500 GPs and more....is that correct???

Response:
We added information about the impact of measurement fluctuation in the results section above table 1. This topic is also cited in the limitations of the discussion section.

3. Page 11 - perhaps I am wrong but I see a clear sex-differential in the estimated mean hours (Table 1) regardless of the type employment albeit more female GPs participated in the study - I am curious to ask if the analysis was stratified by male vs female measurements would there be an change to the results seen in table 2 and figure 2.

Response:
The sample was stratified by gender and employment position, but some subgroups are represented relatively more than others. We agree that we have to account for this misbalance. Therefor we weighted the sample fluctuation by gender and employment position. This is explained in the methods section (analyses). The weighted results are shown in table 1 and figure 2 of the paper.

4. Lastly, page 6 a minor point, in line 157 the authors mention 8 messages per day, so 56 messages per week were scheduled - that is a bit confusing as in a real situation 8 x 5 days so it should be 40 messages per week but I leave it to the authors to improve the line.

Response:
In this case a week includes the weekend days as well. GPs are working during this part of the week too, for example to do administration or out-of-office shifts. We have clarified this sentence in the methods section (data and materials).

Reviewer #2: Report

The paper aims to analyse the precision of the weekly hours worked by GPs in the Netherlands. To do so, more than 1000 GPs were questioned about their activity by sending them one SMS message every three hours during a week. They use real data and simulations to estimate such precision. They come up with two already well known results in statistics:

1 - The larger your sample size, the more precise your estimate will be, and this is true either if the number of GPs is increased or if the frequency of measurement is increased.

2 - The marginal gain in precision decreases as the sample size increases

The paper is clearly-written and the problem is well introduced. Though interesting to read from an application and managerial point of view, at its current state this paper has some methodological flaws:

* Throughout the whole article, and even in the title, the authors repeatedly use Reliability of the estimate when it would be more accurate to use Precision.

Response:

We agree that this is more appropriate and replaced the word ‘reliability’ to ‘precision’ if necessary.

Methods

Data and Materials

* Nothing is said about how the GPs were sampled: it was simple random sampling? Or a stratified sampling scheme was used? Any other sampling design was adopted? Although the authors state (pg. 7, lines 165 to 170) that "The composition of the response group regarding age, gender and position of employment corresponded to a large degree with the composition of the sample that was drawn from the national registration of GPs conducted by NIVEL" it is important for the sampling method used to guarantee the representativeness of the sample.

Response:
We have used a stratified sampling scheme based on gender and employment position. This is clarified in the methods section (data and materials). The population of some of these subgroups was relatively small, e.g. male salaried GPs. For these subgroups we took a larger sample to increase the precision of the results for these GPs. We weighted the data to the gender and employment position of the GPs to account for this misbalance (see also our explanation point 3 of the first reviewer).

Analysis

* Line 200 to 219: "Step 1: determining sample fluctuation". In the formula used to compute the 95% CI, N "represents the total number of measurements of different numbers of GPs". It is not clear to me what this means. Is N the number of GPs? Is N the total number of "messages"? As an example, for the "total" group in Table 1, N = 1051 or 58,856 or something different?

In my understanding, N should be the number of GPs (i.e. 1051 for the "total" group), because all of them are sent a message every 3 hours during the working days. If this is the figure used, please state it, otherwise justify the figure used, for example by citing references.

Response:

As we cited in the methods section (analyses), the N in the formula reflects the number of measurements. For a precise assessment of sample fluctuation we calculated the CI’s on the level of all measurements.

* The paragraph in lines 215-219 is not clear. I don't understand what was done and with which purpose. If the goal is to estimate the average hours worked by GPs in a given week and SMS were actually sent to sampled GPS every 3 hours, what is the point of computing a CI as if they were sent every 2 hours or every half an hour? It is real data or is simulation data? What are they trying to estimate here? Is this just an introduction to Step 2?

Response:

In this paper we showed how the CI’s change by an increasing number of participants. Furthermore, we simulated how the CI changes if more frequent measurements are conducted. We did this in step 1 and step 2, for sample and measurement fluctuation respectively. In the case of sample fluctuation we simulated this by multiplying the data with for example factor 6 for a measurement every half an hour.
* Line 223: why one-tailed 95% CI? Why not two-tailed, if the aim is to obtain "the hours above or below the average value"

Response:

We chose to show the one sided CI. But we agree that two-sided might fit better with the aim stated in the paper. We included a short note that the two-sided CI can be calculated by multiplying the results of the one sided CI by 2.

* Line 220 onwards: Step 2: since simulation is used to estimate the measurement fluctuation, the "population parameter", i.e. the total number of working hours in a week, is fixed (and known for the simulation purposes). Therefore, why not provide an estimate of the mean of number of hours worked. This would be very useful to see whether the SMS approach has any bias in estimating the total number of working hours in a week and, thus, look for methods to correct it.

Response:

We used the 40- hours working week schedule as a fixed parameter, as this is applicable for a large part of the GPs working in the Netherlands (see also our explanation point 1 of the first reviewer).

* Line 251 onwards: Step 3: this is one of the weakest points in the whole paper

The authors estimate the overall variance as the sum of the two variance components: sample fluctuation estimated in Step 1, and measurement fluctuation estimated in Step 2. However, sample fluctuation variance is estimated from real data whereas measurement fluctuation variance is estimated by simulation for a very specific case!! Let me use another possible working week as an example: let's suppose that a fictional GP works Monday to Thursday from 7:30 to 10:30, from 10:30 to 1:30, rests from 1:30 to 4:30 and comes back to work from 4:30 to 7:30pm, Friday works from 7:30 to 1:30, for a total of 42 hours worked in the week. So, the working periods and resting periods coincide exactly with the 3 hour slots. In that case, the variance will be = 0, no matter the frequency! Moreover, different working patterns will yield different measurement fluctuation variability for a given frequency!

Response:

We agree that this is a difficult point. The working week is fictional and the simulations are an attempt to gain insight into this type of fluctuation. More or less measurement fluctuation plays a role in determining the total CI and it depends on the type of working week GPs actually have.
Other working week schedules would have been possible, but we think that this 40-hours working week schedule provides the best insight into this type of fluctuation for the whole group of GPs, as earlier research showed that a large part of the Dutch GPs are working during these moments of a week. The methods section (analyses, step 2) includes a reference to this publication in which the results are reported. We did however conduct additional simulations on other types of working weeks. These results are included in the methods section (analyses, step 2), in the limitations of the discussion section and in the appendix we added a table showing the results of these simulations (see also our explanation point 1 of the first reviewer).

- The overall variability is obtained by adding up the SE2 for the two components. The numerator used to compute the SE for the sample fluctuation has already been discussed. But no formula is presented to compute the SE for the measurement fluctuation. What is the N in this case?

Response:

Simulations are based on probabilities and random models. In this case the N reflects the individual GPs, i.e. 1, 50, 100 etc.

- In any case, the overall variability has to be applied to every Yi (number of working hours worked by one GP in a given week). Therefore, to compute the overall variability, variances and not SE have to be added. The SE will be computed in a second stage, once the overall variance is obtained.

Response:

In our main report we discuss the average working hours and some other descriptive statistics only based on the real data. For example the standard deviation for male self-employed GPs is approximately 10.5.

In this paper we discuss and report the effect of sample size on the CIs accounting for sample and measurement fluctuation. We focus on the CI as this is the main indicator for the precision of the results.

Results

* Line 285: Figure 2 shows …: is this real data or simulated data?

Response:
As explained in the methods section (analyses), both sample fluctuation and measurement fluctuation are based on the equations and simulations on the data. We adapted the text and the footnote of Figure 2.

The context and reason to conduct the simulations is a ‘real’ time-use survey we conducted among over 1,000 Dutch GPs during one year as described in the beginning of the paper.

- Line 289: this is an obvious result, the width of a CI is proportional to the SE, i.e. proportional to the SD/√N, if N goes from 1 to 50, the SE decreases √50=7.07, which is basically the result reported!

- Lines 290-295: again, the result is obvious. From 50 to 100 GP, the sample is doubled, the SE is divided by √2= 1.41, basically the result reported.

Response:

We agree with the reviewer that the results presented are logically derived from applying the standard SE and CI formulas (i.e. the √N element). This point is added in the paper on a numbers of places where we present and interpret the results.

- It is not clear what does the 95% CI measure: (im)precision due to overall variability? Sample variability alone? Measurement variability alone?

Response:

In this paper we show how the total CI (sample and measurement fluctuation) changes if certain design decisions would be made. In the text above table 1 we included information about the impact of measurement fluctuation on the total CI.

Discussion/Conclusions

* Across the whole discussion the precision of the estimates for the total group and the different subgroups are discussed. The results are pretty straightforward and well known, as stated above: the differences in the width of the 95% CI are due to different sample sizes. That's all!!

Response:

We agree the paper include ‘obvious’ results as we apply the standard formulas for calculating CIs. We added this at a number of points in the paper. We believe the added value of this paper
is that we show, in this case of doing time-use research among GPs, how many participants are needed to attain a certain level of precision. In this type of research this cannot be calculated in a straightforward manner as a result of the combination of sample and measurement fluctuation. This appears to be a white spot as we did not find any articles in which the CIs are discussed in detail. The information about the impact on the required participants as a result of different design choices is important information for time use researchers who want to apply this method of time sampling. In the introduction and discussion of the paper we clarify and stress this argument more.

Conclusion

A huge experimental effort seems to have been undertaken in order to get the data. The problem seems relevant but the methodology and the discussion are poor. I strongly recommend the manuscript be revised by convincingly addressing the above issues.

Response:

We adapted the paper according to the points mentioned above.