Author’s response to reviews

Title: Dealing with missing data in the Center for Epidemiologic Studies Depression self-report scale: a study based on the French E3N cohort

Authors:

Noémie Resseguiier (noemie.resseguier@yahoo.fr)
Hélène Verdoux (hverdoux@ch-perrens.fr)
Roch Giorgi (roch.giorgi@univ-amu.fr)
Françoise Clavel-Chapelon (francoise.clavel@igr.fr)
Xavier Paoletti (xavier.paoletti@curie.net)

Version: 2 Date: 11 January 2013

Author’s response to reviews: see over
We are very grateful to the reviewers and the editor for carefully reviewing our work. Detailed comments from Dr I White led us to modify the text, tables and Figures of our manuscript in order to address the important points raised. We followed the sequence of questions and comments.

**Major compulsory revisions**

1. All the results presented assume that interest lies in the mean score or prevalence after dichotomisation. But often we are interested in the association between CES-D and another variable, Y say. In this setting, it is important to include Y in the imputation model for the CES-D items (e.g. Moons et al, 2006): the authors’ conclusion that only the items need to be included would lead to bias in the estimated association. (It could be that the bias would be so small that it doesn’t matter, but the authors need to show this.)

We fully agree with this comment and we have modified the text accordingly. We had used this approach in Resseguier et al, (Epidemiology, 2011) when we illustrated our strategy of imputation under missing not at random scenarios. However as the objective of the present study was to estimate the prevalence of high depressive symptoms, we did not investigate the impact of not including the variable on which to regress in the imputation model.

We have modified the following sentences:

- In the abstract: we have deleted “using an imputation model including CES-D items”
- In the discussion (4th paragraph): we have added “Last, in regression analyses of the association between an outcome and the CES-D scale, the imputation model can be safely adjusted on the CES-D items and on the covariates of interest in order to limit the risk of bias in the association measure [40].”.
- In the conclusion: we have replaced “the available CES-D items responses” with “The imputation model can be restricted to the available CES-D items responses
for score estimate and to the CES-D items and the covariates in regression analyses.”.

2. The methods section needs to be more clearly structured, and to follow the same structure as the results section. I suggest that a separate subsection is made for “Imputation models” to include all relevant details. Also the single imputation approaches must be defined.

Done.

We have grouped methods for multiple imputation and imputation models under the subheading “multiple imputation and imputation models”.

We have modified the presentation of the single imputation approaches.

3. a. It’s not made clear whether the women were asked about item or unit non-response. Surely interest lies in item non-response, so the question should have been e.g. “why did you not answer question 9 but answer other questions?” But then “stressful life event” (table 3) is not a good answer! And the conclusion (p15) “An interview study showed that none of the women contacted declared that they were reluctant to answer the questions of the CES-D scale” only makes sense if they were asked about unit non-response (i.e. about the scale as a whole). Please clarify.

The interviews were standardized but questions were specific neither of item nor of unit non-response. They focused on questions such as “How did the woman perceive the CES-D questionnaire?”, “What did she think about the format of the questionnaire? And its content?”, “How did she evaluate her ability and ease to express her emotions or feelings in different conditions (with her relatives, her friends, and her general practitioner?)”. Potential reasons for missing values which could be associated with depressive symptoms were recorded.
In the Methods section, we have made clear what questions had been asked to the whole sample.

b. The presentation of the results (table 3) is very unclear. It doesn’t help that the table heading refers to an “interview study” which then appears to be made up of a “postal letter” and an “interview study”. But why are reasons given under “postal letter”, when the methods only described reasons being elicited by telephone?

We have tried to clarify this point in the revised manuscript. We now use the word “qualitative study” instead of “interview study” to avoid confusions. Phone interviews only are now identified as the interview study.

Among the 183 responders, 16 women spontaneously included written comments on the filled questionnaires. As shown in table 3, these comments somehow reflected the perception of the women regarding the CES-D questionnaire. All women with missing data were contacted by phone. However, the 11 women who wrote comments and replied to all questions were not contacted by phone. Their written comments were taken as source and analyzed together with oral interviews. We then made clear this distinction in Table 3.

4. Problems with the description and aims of the simulation study, p10:
   a. Were the random MVs created under a MCAR mechanism?
   Yes, random MVs were created under a MCAR mechanism.
   We have modified the sentence “Random MVs were created from the 39,393 complete cases” by adding “under a MCAR mechanism” (Methods section, Simulation study subsection).

   b. “The mean squared error between observed and obtained scores from imputed data was measured”: what is the “obtained score”? Is this mean squared error appropriate? (e.g. Schafer & Graham, 2002: “the goal of a statistical procedure should be to make valid and efficient inferences about a population of interest—not to estimate, predict, or recover missing observations nor to obtain the same results that we would have seen with complete data.”)

   We agree that the goal of multiple imputation is to provide valid inferences for statistical estimates from incomplete data, and that it is not to recover exactly missing values. We
performed a simulation study to compare the predictive properties of the imputation model according to the number of missing values in the CES-D scale. We studied the properties of the imputation model by comparing indicators of predictive accuracy.

In the previously submitted manuscript, we had investigated 2 outcomes: (i) the CES-D score taken as a continuous variable and (ii) the presence of high depressive symptoms of depression taken as a binary variable. For clarity, we have deleted the predictive accuracy evaluation of the binary outcome that used sensitivity, specificity and Cohn’s kappa; we only report the prevalence of high depressive symptoms under various models. We rather focus on the mean and variance of the CES-D score as well as the standard error of the mean CES-D score.

We have modified the text accordingly in the Methods and Results sections.

c. “Sensitivity, specificity and Cohen’s kappa were estimated from observed and obtained presenting / not presenting hDS status”: this needs to be made much clearer. Sensitivity of what?

We have removed these results for the sake of clarity.

d. Supp. material 7 – why are results presented for both 0 MVs and 20 MVs?

We have modified the figure and deleted the results for both 0 MVs and 20 MVs.

e. Would it not be better to use the simulation study to assess how well the different methods (including pro-rating, the current standard method) perform?

We considered that a complete assessment of the competing methods in a simulation study may have been redundant with previous works that illustrated that “standard” single imputation methods are biased. However, we have added the results obtained with the pro-rating method that is the commonly used.

f. I’m not convinced that “A simulation study showed the good predictive properties of the imputation model in this setting.” (p15) is a reasonable conclusion.

We have qualified our statement with “A simulation study showed that multiple imputation performed well, even in the presence of a large amount of MVs and using an imputation model including only CES-D items.”.
5. MNAR imputation, p11: how was this procedure combined with MICE? For example, was the extra parameter included in every imputation model in every cycle? If so, note that van Buuren et al (1999) got contradictory results in this way – they assumed that missing values were 5 units higher than observed, and obtained imputed values that were some 12 units higher than observed.

We proposed a strategy in several steps. The first step consists in the estimation of the parameters for the imputation model, under MAR assumption, using the MICE procedure implemented in the Package R `mice`. Estimation with MICE follows an iterative process in $i$ iterations to obtain convergence. The second step is then to carry out an $i+1$ iteration, and to add an extra parameter in the imputation model for this last iteration. This extra parameter models the OR for the item with missing data between responders and non-responders. The third step is then to impute missing data using the last parameters of the imputation model, under the MNAR scenario thus specified.

As we wanted to control the size effect of the variation imposed in the imputation model, and to avoid a cumulative effect of this variation, we proposed to add it only for the last iteration. More details can be found in *Sensitivity analysis when data are missing not-at-random, Resseguijer et al, Epidemiology, 2011*.

6. How much do the results generalise to other rating scales? I think there is the potential to draw useful generalisations. And would the methods used be applicable to other rating scales?

Our results should be generalizable to other rating scales if missing values in items of these other scales can be considered as ignorable.

Multiple imputation, as well as our strategy of imputation under MNAR scenarios, is applicable to other rating scales. A sensitivity analysis under various scenarios would help researchers to be confident in the ignorable mechanism of missing values they have to deal with.

We have added this comment in the end of the conclusions section: “Methods used in the present study on the CES-D scale can be applicable to other rating scales when dealing with MVs. If MVs can be considered as ignorable, then multiple imputation is recommended.”
Minor compulsory revisions

7. Abstract, conclusions, “In a population of women, missing data mechanism appears to be ignorable” – needs to be qualified / justified.

We have modified this sentence in “Based on the results from both a qualitative study and a sensitivity analysis under various scenarios of missing data mechanism in a population of women, missing data mechanism appears to be ignorable.”

8. p5: “Ignorable or nonignorable missing data mechanism cannot be identified from the data collected.” – well, it can be identified under extra assumptions, e.g. Diggle & Kenward (1994).

We have modified our sentence in “Ignorable or nonignorable missing data mechanism cannot be identified from the data collected, or only in specific contexts under extra assumptions [9]”.

9. Imputation models, p10:
   a. clarify whether the values imputed under a linear regression model were rounded to integers (and if not, whether 16.5 rather than 16 was taken as a cut-off for caseness)

We used predictive mean matching method, that is we computed the predicted value using linear regression, $y^*$, then we selected among observed data the one whose corresponding predicted value was closest to $y^*$.

   b. Please explain why a proportional odds model was not considered

We preferred using polytomous unordered regression model to avoid assuming proportional odds, which was possible due to the large sample size and power. However the linear regression model that makes an even stronger assumption on the relation between the different modalities of the items led to quite similar results. We then expect POM to give the same results too.
10. p12, “suggesting the possibility to efficiently recover from partially missing information” – what does this mean?

We have deleted this part of the sentence.

11. p13, “This interval did not include values observed among complete responders, suggesting that MVs were not MCAR.” – but uncertainty has not been allowed for: in fact the 41.2% figure is not significantly different from the 30.2% figure.

We agree that the 41.2% figure is not statistically different from the 30.2% figure with our sample size. However, imputation of all the missing values successively to 0 and 3 gives prevalence of high depressive symptoms between 41.2% and 64.7% among women who had missing values in the scale, and between 32.2% and 36.6% among the whole sample. We expect the prevalence to lie in these 2 numbers that represent 2 extreme patterns of missing data. As these intervals do not include 30.2%, the point estimate in responders, this suggests that complete case analysis is biased and hence that MCAR mechanism is unlikely. We have replaced the sentence by the following one: “Imputation of all the MVs successively to 0 and 3 gave prevalence of hDS between 41.2% and 64.7% among women who had MVs in the scale, and between 32.2% and 36.6% among the whole sample (Table 2). These intervals did not include the point estimate based on complete data only, suggesting that MVs were not MCAR.”.

p13, “None of the women contacted refused to participate in the interview.” – so why does table 3 report interview results for only 29 of 34 women? Presumably 5 couldn’t be contacted? This should be stated.

Among the 34 women who had missing values, five of them wrote directly a comment on the questionnaire and we retained this information as a source (consistent with what they said by phone).
12. p17, under MNAR, “the absolute difference in prevalence was … less than 1% in the subsample of women with up to ten MVs”: this is inappropriate and should be deleted, since under MNAR it is wrong to exclude women with > 10 MVs.

We fully agree and we have deleted this part of the sentence.

13. Table 4:

a. How where the SDs computed? I guess Rubin’s rules were used (i.e. the SD was computed in each imputed data set and then averaged)?

We used Rubin’s rules to compute the SDS. We computed the variances in each imputed dataset, we then averaged them, and finally computed the SDs.

b. It would be helpful to report the standard error of the mean (NB this cannot be derived from the SD and N once we have incomplete data), since comparisons of the standard error between methods will show how much information was in the incomplete records.

We have computed and added the standard error of the mean for the various methods.

14. Table 5: the definition of the theta parameter is not clear, nor is the phrase “the modality of interest” in the footnote. I guess theta has 3 values for the 3 contrasts of a response level vs. baseline response level? But how were the particular values of theta chosen? Why is it sensible to have the 3 values increasing?

We have replaced the phrase in the footnote by: “Vector of parameters for the MNAR scenario, each parameter corresponding to the odds ratio expressing the excess risk to respond one of the modalities of interest (i.e. an item score of one, two or three) compared to the reference (i.e. an item score of zero), in subjects with MV compared to subjects without MV.”.

Choice of an adequate scenario for sensitivity analysis is a delicate issue. We assumed that nonresponders were more likely to have high scores than responders (theta>1). For Scenario 1 and Scenario 2, the ORs in the imputation models between responders and non responders for
a given item were (1.2, 1.5 and 2.0 for the modalities 1, 2 and 3, respectively). These values are in line with the observed odds ratios between high depressive symptoms and various covariates. For scenario 2, we additionally supposed a stronger OR for positive items. For Scenario 3 and Scenario 4, magnitudes were higher.

Higher values for the theta parameter when considering positive items, as well as increasing values according to the modalities 1, 2 and 3 are in line with the results illustrated by Olino TM, et al., 2012 in a study using item response theory methods that showed higher discrimination parameters values for positive items and showed that the subjects are more difficulty to rate higher values.

15. Typos:
   a. p4, l9: Factor (not factorial) analysis
   b. p5, l-1, “two different imputation models”
   c. p10, l2: “mixed-type covariates”; l3, delete “an”.
   d. p12, “the rather unidimensional structure”
   e. p19, “CES-D item responses”

We have corrected these typos.

Discretionary revisions

16. p4: presumably the items are scored 0/1/2/3?

Yes, the items are scored 0/1/2/3. We have added this information in the first paragraph of the background section “each item being scored 0/1/2/3”.

17. p18: I think the comparison with other cohorts could be deleted without loss.

We have preferred to keep this part of the discussion in order to get a better understanding of the specific features of the E3N cohort.

18. Table 1 might be more concisely presented as a cumulative frequency graph.

We have deleted Table 1, and we have replaced it with a cumulative frequency graph.