PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (see an example) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below. Some articles will have been accepted based in part or entirely on reviews undertaken for other BMJ Group journals. These will be reproduced where possible.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Mortality by education level at late-adult ages in Turin: a survival analysis using frailty models with period and cohort approaches.
AUTHORS	Zarulli, Virginia; Marinacci, Chiara; Costa, Giuseppe; Caselli, Graziella

VERSION 1 - REVIEW

REVIEWER	Prof. Dr. Andreas Wienke Institute for Medical Epidemiology, Biostatistics and Informatics University Halle-Wittenberg There are no competing interests
REVIEW RETURNED	14-Mar-2013

THE STUDY	page 11, line 13ff I do not agree with the statement that the better fit of the frailty models strengthens the validity of the selection hypothesis. You just see better fit with a more complex model, but to validate the idea of selection you completely different study designs, see for example discussions in Experimental Gerontology (stress experiments).
GENERAL COMMENTS	This is a paper dealing with the effect of univariate parametric frailty models. The presentation of the material is well organized and clear, there are no typo's or errors. I have a few comments: comments: page 1, line 20: it is unclear why Graziella Caselli is mentioned with her title of a Honorary Professor and the other authors not. I think in BMJ Open it is not common to give titles. page 6 it is unclear why table 1 is included in the text whereas the other tables are shifted to the appendix page 6, line 23 I cannot find tables A2 and A3 page 6, line 41 " a slight but not significant reduction" how did you perform the test for this statement? Why did you test not the difference in the other cases where the parameter estimates of the fixed effects increases after introduction of the frailty? This is selective testing which is not useful! page 11, line 13ff I do not agree with the statement that the better fit of the frailty models strengthens the validity of the selection hypothesis. You just see better fit with a more complex model, but to validate the idea of selection you completely different study designs, see for example discussions in Experimental Gerontology (stress experiments). page 13ff please check all reference, Journal titles should be with capital letters (for example Statistics in Medicine instead of Statistics in medicine, ref. 20, 24,25,35,39…) What is ref. 22? A book? Publisher? Appendix A: I do not see what is new in the appendix. Everything is

	well known and can be found in text books about frailty models. Appendix B: I do not understand what is the sense of the exponential model? I did not found it in the text, where the Gompertz model was used. What is the additional benefit of the exponential model compared to the Gompertz model? page 21 covariates A,B,C are a,b,c in the formulas page 21 line 16 sigma^2 for a non-frailty model should be shifted one line above (parameter c) page 21, line 17 “cohort” instead of numbers in the table? page 21 table C1 and C2 estimates and CI with 0.000 should be rescaled to provide reasonable information Which software was used? Stata? Was the corrected likelihood for left truncated data used suggested by van den Berg and Drepper (2011)?

REVIEWER	Anton E. Kunst Department of Public Health Academic Medical Centre (AMC) University of Amsterdam
REVIEW RETURNED	29-Mar-2013

THE STUDY	The educational differences demonstrated in this paper appear to be sensitive to control for either period or cohort. Compare the estimates for education in table C1 to those in C2, for ‘models without frailty’ (or see Figure 2): the RR’s vary substantially (e.g. 1.25 towards 1.20). One would expect the authors to be able to adequately and unambiguously control for the effect of ‘time’ in the measurement of educational differences in mortality. For readers who are not familiar with frailty models, it is not possible to understand, just from formula presented in the appendix, what frailty models in essence are. Furthermore, it is difficult to understand how such models could possibly influence estimates of educational differences in mortality. A way to intuitively understand such an influence would be that the measure of ‘frailty’ was related to educational level. However, it is not clear whether and how this happens.
RESULTS & CONCLUSIONS	While the authors state repeatedly that models with frailty have a better fit, the differences in AIC in Table seem extremely small, and they are not consistently better or worse for models with the frailty parameters. The text should be corrected at this point, including the conclusions the very end. In a similar way, the authors overstate the evidence, presented in Figure 2, to support their claims for a better performance of frailty models. Larger educational inequalities are observed in only 2 out of 4 cases (one of which cannot be demonstrated with statistical significance). Moreover, for ‘period’, there are problems in comparing the results with and without frailty models (see p 6/7). The comparison that can be made with greater accuracy for ‘cohort’ shows contrasting results for men and women. The first objective of the paper, which is about ‘convergence’, is not addressed in the empirical analyses. For this, the authors should have used frailty models to assess educational differences in mortality AT DIFFERENT AGES. The authors should be careful to present their results as evidence to

	suggest observational studies result in ’underreporting’ or ‘bias’ of inequalities in mortality. Please note that observational studies simply observed the highly important fact that, at older ages, age-specific rates vary between educational levels by a certain amount. The role of frailty model is not to ‘correct’ a presumed bias in these observations, but to help to understand them with reference to underlying mechanisms.
GENERAL COMMENTS	I could not assess the complex statistics. This should be done by a statistician.

VERSION 1 – AUTHOR RESPONSE

Andreas Wienke.
There are no competing interests.
This is a paper dealing with the effect of univariate parametric frailty models. The presentation of the material is well organized and clear, there are no typo’s or errors. I have a few comments:

COMMENT: page 1, line 20: it is unclear why Graziella Caselli is mentioned with her title of a Honorary Professor and the other authors not. I think in BMJ Open it is not common to give titles.
ANSWER: The affiliation has been changed.

COMMENT: page 6 it is unclear why table 1 is included in the text whereas the other tables are shifted to the appendix.
ANSWER: We had shifted the other tables to the appendix to keep the main text streamlined, as these tables are quite complex and big. Moreover part of the results contained in these tables is illustrated by figure 2, which is included in the text. But we agree with the reviewer that, for the sake of consistency, all the tables should be included in the same section.

COMMENT: page 6, line 23 I cannot find tables A2 and A3.
ANSWER:Tables A2 and A3 were, in reality, tables C2 and C3 in appendix C. We apologize for the wrong naming. The tables are now Table 2 and table 3 in the main text.

COMMENT: page 6, line 41 “ a slight but not significant reduction” how did you perform the test for this statement? Why did you test not the difference in the other cases where the parameter estimates of the fixed effects increases after introduction of the frailty? This is selective testing which is not useful!
ANSWER: The estimates were compared in all the cases by using the same procedure of looking at whether the confidence intervals overlap or not. We agree with the reviewer that the expression “a slight but not significant reduction” is unfortunate and misleading, implying that in only this case a test was performed, while in the other cases not. The sentence has been modified.

COMMENT: page 11, line 13ff I do not agree with the statement that the better fit of the frailty models strengthens the validity of the selection hypothesis. You just see better fit with a more complex model, but to validate the idea of selection you completely different study designs, see for example discussions in Experimental Gerontology (stress experiments).
ANSWER: We agree with the reviewer that different study designs are needed to unequivocally validate the selection hypothesis.
The AIC method for model selection does not dismiss the validity of the model with the lowest AIC value but compares models that are all well-supported hypotheses and indicates the model that is likely to approximate better the (unknown) reality, under the assumption that model selection is a way of approximating, rather than identifying, full reality (Burnham, Andreson 2003 – Multimodel Inference: Understanding AIC and BIC in Model Selection). The best AIC shows just that a more complex model better fits the data. We agree with the reviewer that this is not an unequivocal proof of the selection

hypothesis. However, it points to the possibility that the data are better described by this hypothesis. We modified the text in order to clarify better this point (page 11, last paragraph).

COMMENT: Please check all reference, Journal titles should be with capital letters (for example Statistics in Medicine instead of Statistics in medicine, ref. 20, 24,25,35,39…) What is ref. 22? A book? Publisher?
ANSWER: The references have been checked and all the Journals are now written with capital letter. Reference 22 has been specified better.

COMMENT: Appendix A: I do not see what is new in the appendix. Everything is well known and can be found in text books about frailty models.
ANSWER: We agree with the reviewer that formulas and concepts in appendix A are not new. However, the appendix has the function to summarize the main formulas and the newest results about frailty models (like the one referred to the corrected likelihood function for left truncated data) in a unique section. We think that such a summarizing section could be useful to the readers who are interested in knowing more technical details but are not familiar with the topic. This would help them saving time in the search of a very extensive literature such as that about frailty models.

COMMENT: Appendix B: I do not understand what is the sense of the exponential model? I did not found it in the text, where the Gompertz model was used. What is the additional benefit of the exponential model compared to the Gompertz model?
ANSWER: The exponential baseline hazard does not change with age. This allowed including the age as a covariate and to have it interacted with the covariate for education level. This was done to investigate whether there is a statistically detectable convergence of hazards at old ages by education group, by testing whether there is a significant interaction between the variables education and age. The single parameter baseline hazard was modulated by the covariate for the age groups. The identity between an exponential hazard modulated by an age covariate and the Gompertz model makes such exponential models appropriate for human adult mortality data. More detailed explanations have been added to appendix B.

COMMENT: page 21 covariates A,B,C are a,b,c in the formula.
ANSWER: The covariates in the table have been changed to small letters to make them consistent with the formula.

COMMENT: page 21 line 16 sigma^2 for a non-frailty model should be shifted one line above (parameter c).
ANSWER: The misplacement of the line for the sigma^2 parameter has been corrected.

COMMENT: page 21, line 17 “cohort” instead of numbers in the table?
ANSWER: The word “cohort” has been replaced by its numerical estimate.

COMMENT: page 21 table C1 and C2 estimates and CI with 0.000 should be rescaled to provide reasonable information.
ANSWER: The estimates have been rescaled.

COMMENT: Which software was used? Stata?
ANSWER: All computations were performed with the software R. This is specified in the last paragraph of the section Data and Methods (page 5).

COMMENT: Was the corrected likelihood for left truncated data used suggested by van den Berg and Drepper (2011)?
ANSWER: Yes, the likelihood for left truncated data was suggested by van den Berg and Drepper

(2011).

Anton Kunst.
I have no competing interests.

COMMENT: The educational differences demonstrated in this paper appear to be sensitive to control for either period or cohort. Compare the estimates for education in table C1 to those in C2, for 'models without frailty' (or see Figure 2): the RR's vary substantially (e.g. 1.25 towards 1.20). One would expect the authors to be able to adequately and unambiguously control for the effect of 'time' in the measurement of educational differences in mortality.
ANSWER: Time comprises of three simultaneous dimensions: age, period and cohort. To unambiguously control for the effect of time, these dimensions must be taken into account simultaneously. However, when simultaneously controlling for them, the literature on Age-Period-Cohort (APC) models has shown that serious identification problems occur, due to linear dependence between all these three dimensions.
Therefore, we adopted two approaches for the control of it, an age-cohort approach and an age-period approach. We are aware that they represent two different dimensions of time, but they also represent two of the most common approaches taken when addressing demographic questions related to long observation periods and to the mortality improvement occurring during these periods. Moreover, as the two approaches represent each only a partial dimension of the effect of time, it was not expected to find the same or very similar estimates because in the two cases the reference categories, to which the coefficients refer to, are different and represent very different "entities". In the first case, the age-mortality profile of a single cohort taken as reference, in the other case a period of calendar years whose age mortality profile is determined by the overlapping of individuals in different age groups but belonging to many different cohorts.
A paragraph has been added in the Data and Methods section (page 4, paragraph 6) to discuss this issue.

COMMENT: For readers who are not familiar with frailty models, it is not possible to understand, just from formula presented in the appendix, what frailty models in essence are. Furthermore, it is difficult to understand how such models could possibly influence estimates of educational differences in mortality. A way to intuitively understand such an influence would be that the measure of 'frailty' was related to educational level. However, it is not clear whether and how this happens.
ANSWER: We agree with the reviewer that the original version of the appendix are not enough to present clearly what frailty models are to readers that are not familiar with them. The appendix has been expanded in order to include more extensive explanations.

COMMENT: While the authors state repeatedly that models with frailty have a better fit, the differences in AIC in Table seem extremely small, and they are not consistently better or worse for models with the frailty parameters. The text should be corrected at this point, including the conclusions the very end.
ANSWER: We agree with the reviewer that with quite large values such as those in the analysis, a difference of around 270 might seem trivial. However, as discussed in (Burnham, Andreson 2003 – Multimodel Inference: Understanding AIC and BIC in Model Selection), the AIC values themselves are not interpretable as they are much affected by sample size and arbitrary constants such the likelihood value, but they need to be interpreted in terms of their difference. So if ∆_i is the difference between model i and the model with the lowest AIC, "the larger the ∆_i, the less plausible is fitted model i as being the best approximating model in the candidate set. […] in assessing the relative merits of models in the set: models having a ∆_i≤2 have substantial support (evidence), those in which 4≤∆_i≤7 have considerably less support, and models having ∆_i>10 have essentially no support " (Burnham,

Andreson 2003, p.271).

COMMENT: In a similar way, the authors overstate the evidence, presented in Figure 2, to support their claims for a better performance of frailty models. Larger educational inequalities are observed in only 2 out of 4 cases (one of which cannot be demonstrated with statistical significance). Moreover, for ‘period’, there are problems in comparing the results with and without frailty models (see p 6/7). The comparison that can be made with greater accuracy for ‘cohort’ shows contrasting results for men and women.

ANSWER: We would like to point out that, when frailty is taken into account, larger educational inequalities between high group and medium and low groups are observed in the majority of the cases. Beside the case of men in the age-cohort approach model, also for both the cases of men and women in the age-period approach, when frailty is controlled for, the rate ratios of medium and low education groups compared to the high groups lie in a higher confidence region than in the models without frailty, thus pointing to the possibility of having larger educational inequalities between the high group and the rest of the population. What was less clear in this case, was the difference between medium and low rate ratios estimated among men.

However, we agree with the reviewer that this evidence is less strong, particularly because it was not possible to compare the models via the AIC, due to the peculiar statistical procedure needed to estimate these models. A sentence has been added to the discussion to give proper emphasis to this point (page 9).

We also agree with the reviewer that it is important to notice that in one case, the age-cohort approach, men and women show contrasting results. Possible explanations for this pattern have been discussed in the discussion (page 10, first paragraph).

COMMENT: The first objective of the paper, which is about ‘convergence’, is not addressed in the empirical analyses. For this, the authors should have used frailty models to assess educational differences in mortality AT DIFFERENT AGES.

ANSWER: We agree with the reviewer that the issue of convergence was not sufficiently discussed among the results. A paragraph in the Results section (page 8) has been added. As specified in the now extended appendix about frailty models, the issue of the convergence of hazards is addressed through the estimate of the variance. Converging hazards are the result of the effect of selection on the population hazards, due to how much variance of unobserved frailty is present in the population at the initial age of observation (or beginning of the follow up). A detected null variance at the initial age would lead to no convergence, a detected (positive) variance, on the contrary, leads to such phenomenon. Moreover, the bigger the variance the stronger the convergence is. The explorative analysis had showed that the hazards converged more among women than among men. All the models of the regression analyses have found a higher variance among women than among men, consistently with what “predicted” by the frailty models framework.

COMMENT: The authors should be careful to present their results as evidence to suggest observational studies result in ’underreporting’ or ‘bias’ of inequalities in mortality. Please note that observational studies simply observed the highly important fact that, at older ages, age-specific rates vary between educational levels by a certain amount. The role of frailty model is not to ‘correct’ a presumed bias in these observations, but to help to understand them with reference to underlying mechanisms.

ANSWER: The results of the paper don’t intend to suggest that all observational studies underreport inequalities in mortality. The paper rather wants to warn about the risk of obtaining underestimated and biased differentials in mortality, when performing survival analysis studies using longitudinal data series and study designs. We have modified a sentence of the introduction in order to make it clearer that the risk of bias refers to this kind of analysis and not to all observational studies.

We believe that frailty models have both the roles of helping to understand the underlying mechanisms of selective mortality (this has been discussed in the first paragraph of the Introduction and in the last paragraph of the Conclusion) and to correct the estimates when performing a particular type of analysis, and that both roles should be mentioned. A wide statistical literature has pointed out how not-including the unobserved heterogeneity component in regression models can lead to biased estimates. Therefore, frailty models are not only a theoretical framework. They are also a set of statistical methods, namely random effect models where the random term represents the unobserved frailty component, that allow correcting for the bias caused by an ignored heterogeneity component in a specific study design: a longitudinal (or cohort) analysis that uses survival analysis methods.

COMMENT: I could not assess the complex statistics. This should be done by a statistician.

VERSION 2 – REVIEW

REVIEWER	Andreas Wienke Institute of Medical Epidemiology, Biostatistics and Informatics University Halle
REVIEW RETURNED	07-May-2013

THE STUDY	I would like to mention one additional point regarding the problem of the three dimensions age, period, and cohort (page 4). The statement that so called APC models are not identifiable is not completely true and based on older references (37-39). There exists recent work on this area with newer results. Making a few additional assumptions APC models become identifiable. I recommend the book "Age-Period-Cohort Analysis" by Yang and Land (2013), which provides software to run such models.

REVIEWER	Anton E. Kunst Department of Public Health Academic Medical Centre (AMC) University of Amsterdam
REVIEW RETURNED	16-May-2013

RESULTS & CONCLUSIONS	Though the authors have responded to my comments, I feel that these responses have not all been satisfactory. Comment 1. Control for time. When the issue is to control for the time (age, period, cohort) as a confounder, the challenge is to removing its confounding effect as much as possible, and to avoid ‘residual confounding’. This requires fitting one model that captures as much as confounding effects of age, period and cohort. The ‘identification problem’ is not really a problem here, because the aim is not to separate cohort from period effects. But if the authors are not happy with including both exchangeable parameters of age, period and cohort, they could have introduced some restrictions to the model (e.g. continuous parameters for cohort and period), in order to avoid the identification problem. My key recommendation is thus that, instead of applying two models with each some residual confounding, the authors apply just one model with as complete control as possible. At the very least, I would recommend that the authors avoid using “the model with period improvement” because this model is highly complex and it resulted in very imprecise estimates (wide 95 CI) without having any clear advantage. Comment 2. More information on frailty modes. This response is okay.

	Comment 3. Interpretation of differences in AIC in table 1. I understand the explanation given by the authors, which is agreeable. The only thing that remains problematic is the interpretation to table 1 as given in the text. Please note that, of the 4 comparisons that can be made, only 2 comparisons support the statement that including frailty yields a better fit. In two cases, the difference in AIC is only 2 units, which is null, also according to Burnham, Andreson 2003. Comment 4. Overstating the evidence presented in Figure 2? I may agree with the response of the authors. However, as the methodological value of the “the model with period improvement” seems unclear, while the disadvantages of this model are obvious, the authors may give much greater weight to the models on “cohort improvement”. This re-emphasis may result in a further modification of the text. Comment 5. Evidence related to the first objective of the paper, which is about ‘convergence’. The response on this point is not satisfactory. If the issue of ‘convergence’ is presented as one of the two key objectives, the reader would expect that much of the Results section is dedicated to addressing this objective. However, the authors include only one paragraph at the end of the Results section, and the evidence presented on this issue is highly abstract and – to me as lay person – incomprehensible. To make things easy, my suggestion is to simply reformulate the objectives of this paper. Comment 6. Evidence to suggest ’underreporting’ or ‘bias’ in studies inequalities in mortality? The authors responded that this problem does not apply to “all observational studies underreport inequalities in mortality” but only to “survival analysis studies using longitudinal data series and study designs”. This sounds reasonable. It is however unclear what type of studies the authors have in mind. In a way, every ‘mortality’ analysis is a ‘survival’ analysis, and mortality is commonly studied using a longitudinal design. So the question remains: which studies may be biased and which are not? To be more specific: are references 1 to 3 biased according to their views?
GENERAL COMMENTS	I think that the paper can be accepted in the end, but that the authors should be pressurised to take my comments into account once more. I hope that you understand that I cannot give the time to review this complex paper a third time.

VERSION 2 – AUTHOR RESPONSE

Reviewer: Andreas Wienke

COMMENT: I would like to mention one additional point regarding the problem of the three dimensions age, period, and cohort (page 4). The statement that so called APC models are not identifiable is not completely true and based on older references (37-39). There exists recent work on this area with newer results. Making a few additional assumptions APC models become identifiable. I recommend the book "Age-Period-Cohort Analysis" by Yang and Land (2013), which provides software to run such models.
ANSWER: We would like to thank the reviewer for informing us about this very recent publication containing the newest results. We have included this reference in the paper and modified the text in

order to mention briefly the methodological enhancement.

Reviewer: Anton Kunst

Though the authors have responded to my comments, I feel that these responses have not all been satisfactory.

COMMNET 1: Control for time. When the issue is to control for the time (age, period, cohort) as a confounder, the challenge is to removing its confounding effect as much as possible, and to avoid 'residual confounding'. This requires fitting one model that captures as much as confounding effects of age, period and cohort. The 'identification problem' is not really a problem here, because the aim is not to separate cohort from period effects. But if the authors are not happy with including both exchangeable parameters of age, period and cohort, they could have introduced some restrictions to the model (e.g. continuous parameters for cohort and period), in order to avoid the identification problem. My key recommendation is thus that, instead of applying two models with each some residual confounding, the authors apply just one model with as complete control as possible. At the very least, I would recommend that the authors avoid using "the model with period improvement" because this model is highly complex and it resulted in very imprecise estimates (wide 95 CI) without having any clear advantage.
COMMENT 4: Overstating the evidence presented in Figure 2? I may agree with the response of the authors. However, as the methodological value of the "the model with period improvement" seems unclear, while the disadvantages of this model are obvious, the authors may give much greater weight to the models on "cohort improvement". This re-emphasis may result in a further modification of the text.
ANSWER: We believe that the period improvement model has the valuable advantage of taking care of the improvement of mortality conditions in the most correct way possible, as this is a phenomenon that takes place mainly on a period basis and not on a cohort basis (for example, penicillin or other medical breakthrough discoveries happen in a specific calendar period and start affecting all the cohorts lucky enough to pass through that period, but each of them is affected at very different ages). In the specific case of our analysis, for the purpose of controlling for unobserved heterogeneity it was necessary the application of the shared frailty models and this required, indeed, a very complex model and alternative methods of estimations (bootstrapping). However, we believe that the high degree of complexity and "heaviness" of the computation do not mean that the investigator has to give up the possibility to try to estimate the model.
Regarding the imprecision of the estimates, we would like to point out that in this particular case, bootstrapping methods were used. Bootstrapping provides, by definition, less precise estimates and wider confidence regions (that are not normal confidence intervals) than normal regression models. Nevertheless, although estimates are indeed less precise, we believe that the information coming from these estimation procedures should not be dismissed.
However, as every result must be considered in the light of the estimation procedure used to obtain it, we agree with the reviewer that such results might be given a smaller weight compared to those, more precise, obtained with the cohort improvement model. Therefore, we have furthermore modified the text (4th paragraph at page 11).

COMMENT 2: More information on frailty modes. This response is okay.

COMMENT 3: Interpretation of differences in AIC in table 1. I understand the explanation given by the authors, which is agreeable. The only thing that remains problematic is the interpretation to table 1 as given in the text. Please note that, of the 4 comparisons that can be made, only 2 comparisons support the statement that including frailty yields a better fit. In two cases, the difference in AIC is only

2 units, which is null, also according to Burnham, Andreson 2003.
ANSWER: The comparison in Table 1 is aimed to compare 4 different models for the mortality hazard. Two models don’t include the unobserved heterogeneity component (and therefore, they can be considered also as baseline hazards), two models incorporate this component. The results show that the best baseline hazard is Gompertz for the male data and Makeham for the female data. When the models using the best baselines incorporate also the unobserved frailty component, the fit significantly improves (AIC differences of 183 and 38).
The two cases of almost null AIC difference (a difference of 2) refer to the not optimal baseline hazards (Makeham for the men and Gompertz for the women). When adding the extra parameter, complexity is added to complexity, but the basic setting remains, however, suboptimal, as the baselines are not the optimal ones. The inclusion of the frailty component alone does not necessarily yield a better fit if the functional form of the baseline hazard is poorly fitting the data. To better understand this point, it must be bared in mind that the AIC comparison does not perform a significance testing where the added parameter is tested for its significance, like a likelihood ratio test between nested models, but compares the different models on the level of their overall functional form.
To avoid misunderstandings we have modified table 1 by removing the comparison frailty/no frailty in the cases of not optimal baselines.

COMMENT 5: Evidence related to the first objective of the paper, which is about ‘convergence’. The response on this point is not satisfactory. If the issue of ‘convergence’ is presented as one of the two key objectives, the reader would expect that much of the Results section is dedicated to addressing this objective. However, the authors include only one paragraph at the end of the Results section, and the evidence presented on this issue is highly abstract and – to me as lay person – incomprehensible. To make things easy, my suggestion is to simply reformulate the objectives of this paper.
ANSWER: We have followed the reviewer’s suggestion. We have reformulated the objectives of the paper and we have modified/deleted some parts of the text to give less emphasis to the convergence issue.

COMMENT 6. Evidence to suggest ’underreporting’ or ‘bias’ in studies inequalities in mortality? The authors responded that this problem does not apply to “all observational studies underreport inequalities in mortality” but only to “survival analysis studies using longitudinal data series and study designs”. This sounds reasonable. It is however unclear what type of studies the authors have in mind. In a way, every ‘mortality’ analysis is a ‘survival’ analysis, and mortality is commonly studied using a longitudinal design. So the question remains: which studies may be biased and which are not? To be more specific: are references 1 to 3 biased according to their views?
ANSWER: The types of studies we refer to are studies that use the specific group of statistical techniques called “survival analysis”, that is, analysis that use the specific methodology of duration dependence models. Even if the statistical terminology might sound confusing, survival analysis studies are only this kind of studies, and not all mortality studies.
Some mortality studies are expressly cross sectional. Others are longitudinal, but among them, only those using survival analysis techniques are survival analysis studies. Having longitudinal data or a longitudinal design per se does not qualify a study as a study of survival analysis (in the technical and statistical sense). If the data are longitudinal but are grouped in age categories and then logistic and/or poisson regressions are applied, another kind of analysis is being done and not a survival analysis. In this kind of analyses, in fact, the outcome of interest (dependent variable) is dead/alive, and the duration of the process enters the analysis as an independent variable represented by the age group categories. In studies of survival analysis, on the contrary, the duration of the process (time until death) becomes the dependent variable.
Therefore, the risk of bias does not apply to the references mentioned by the reviewer, as these studies did not apply the body of techniques of duration dependence models.
In order to avoid misunderstandings and confusion, we have modified the second sentence of the last

paragraph at page 2, by adding the specification: “…not controlling for it, in models of survival analysis…”.

FINAL COMMENT: I think that the paper can be accepted in the end, but that the authors should be pressurised to take my comments into account once more.