Author’s response to reviews

Title: Measurement error in a multi-level analysis of air pollution and health: a simulation study.

Authors:

Barbara Butland (b.butland@sgul.ac.uk)
Evangelia Samoli (esamoli@med.uoa.gr)
Richard Atkinson (atkinson@sgul.ac.uk)
Benjamin Barratt (benjamin.barratt@kcl.ac.uk)
Klea Katsouyanni (klea.katsouyanni@kcl.ac.uk)

Version: 1 Date: 26 Nov 2018

Author’s response to reviews:

Response to Reviewers

We thank the editor and reviewer for their careful consideration of our manuscript. A point by point response to comments is provided below:

Editor

1) Your paper is of interest for the epidemiological research community. Please address the reviewer's concerns. Additionally explain why you include both NO2 and PM10 where the scenarios are exactly the same (only the betas differ). So it is unsurprising that the trends observed for both pollutants are the same. Consider to move one pollutant to the supplement, reword your observations.

Response: The NO2 and PM10 simulations differ in various ways. As you point out the betas differ which means that the “true” outcome distributions will differ between pollutants. In addition distributions of “true” NO2 and PM10 data will not be the same, as they are informed by an analysis of their respective monitoring data. For example when simulating “true” NO2 data we incorporate far more variability both temporally and spatially than for PM10 (see paragraph 2 of the methods) and a different pattern of spatial covariance. Different underlying distributions of “true” data will then result in different distributions of “pseudo” modelled data even when the scenarios in terms of the variance ratio and correlation coefficient are the same. Given suggestions in two previous simulation studies [7,12] (see discussion , page 16, second paragraph) that the adverse effects of measurement error on health effect estimation may be moderated if there is high spatial correlation in the underlying “true” exposure surface, we felt it was important that we did not confine ourselves to just one underlying “true” pollutant
distribution. While trends in % bias are similar for both pollutants, results relating to coverage probability and short-term exposure differ (see Additional File3, Fig 3.1). On reflection we therefore feel that both pollutants should feature within the main manuscript and that the comparison between the two is a useful one.

2) To improve applicability also discuss more which of the evaluated combinations are more likely. Is the setting that showed bias away from the null expected in practice? With a very high correlation, large differences in variances may not be that likely?

Response: This is an important point, but is nevertheless difficult to answer as correlation coefficients and variance ratios are not routinely reported together by model monitor validation studies. A recent study which did report both metrics was one involving pollutant outputs from the EMEP4UK dispersion model. [Butland et al, Open Heart 2016;3:e000429, Supplementary Table 1] The most extreme combinations were observed for rural background PM2.5 (variance ratio=2.68, correlation coefficient=0.73), urban background O3 (0.91, 0.76), rural background PM10 (2.56, 0.47), and urban background PM2.5 (0.61,0.69). However this is only one example and relates to a specific dispersion model and location (i.e. the UK), so how typical the findings are is difficult to assess. Nevertheless as now mentioned (bottom of page 13 / top of page 14), the high correlation low variance ratio combination which produced non-trivial positive bias “may only occur in practice if there is a lack of independence between the Berkson component of measurement error and the modelled data. [9-10]”.

3) In the abstract and method I also did not see how you distinguished / what you assumed type of error.

Response: In the abstract we have gone for simplicity by referring to measurement error as a whole rather than specific types. In the methods section we explain how we incorporate both types of error into our “pseudo” modelled data using a combination of variance ratios and correlation coefficients. As noted (page 9, paragraph 3) the source of our methodology is a paper by Reeves et al [25], which indicated that a second order regression equation [26] expressing “true” data (X) as a linear predictor of model data (Z) could be written to include both classical and Berkson error terms. The use of our methodology in the context of an epidemiological time-series analysis is explained in a previous paper [5] and the extension to a multi-level model and the re-parameterisation in terms of variance ratios and correlation coefficients is explained in Supplementary material 2. (See also response to point 1.1 below).

Reviewer #1:

In this interesting study the authors set up simulations to quantify bias and loss of power in the relationship between air pollution and health induced by measurement error in the exposure. The study has some important strength points, including: a) the joint investigation of short-term and long-term effects; b) the attempt to make explicit all the modelling assumptions about the spatial and temporal components of observed data as well as of the errors in the modelled data.
Given the complexity of the topic (measurement error theory) and of the algebra applied, I would suggest the authors to be more clear on the following aspects.

1.1. Step 4. This is the most important part of the paper, as it introduces Berkson and Classical errors in the simulated data, and it is the most difficult to follow, even in the Appendix 2. In particular, it is not straightforward to understand, in the different addenda of the formulas, which are inherent to Classical and which to Berkson error.

Response: The separation of equation (5) into a part representing classical error and a part representing Berkson error is not possible as we are not simply adding these two errors to “true” data to produce “pseudo” model data. Rather our aim is to introduce both error types by simulating “pseudo” model data which has on average a pre-specified correlation with the “true” data and a pre-specified variance ratio. Given the variance of the “true” data (Var(X)), the importance of the correlation coefficient (\( \tau \)) and the variance ratio (\( \lambda \)) is clear simply from a consideration of the standard formula for total measurement error between model (Z) and true (X) data i.e. \( \text{Var}(X-Z) = \text{Var}(X) + \text{Var}(Z) - 2\text{COV}(X,Z) = \text{Var}(X)(1+\lambda-2\tau\sqrt{\lambda}). \)

We now include the following text within the discussion (bottom of page 12 / top of page 13):

“The aim of our methodology was to introduce measurement error of both types (i.e. classical / classical-like and Berkson / Berkson-like) by simulating “pseudo” model data which had on average a pre-specified correlation with the “true” data and a pre-specified variance ratio both spatially and temporally. The importance of the correlation coefficient (\( \tau \)) and the variance ratio (\( \lambda \)) is clear simply from a consideration of the standard formula for total measurement error between model (Z) and true (X) data i.e.

\[
\text{Var}(X-Z) = \text{Var}(X) + \text{Var}(Z) - 2\text{COV}(X,Z) = \text{Var}(X)(1+\lambda-2\tau\sqrt{\lambda}).
\]

We have also re-written equation 5 on page 9, to better reflect the separation of temporal and spatial components and how we simulated the “pseudo” modelled data.

1.2. The authors introduce correlation terms between true and model data (both spatial and temporal), as well as ratios between their variances (spatial and temporal). According to Armstrong (ref [1]), there is an equivalence between squared correlation between true-model data and ratio of their variances. In this paper the authors treat the two as separate terms (alphas and lambdas). This should be further discussed.

Response: The Armstrong paper notes this equivalence in the context of additive classical measurement error i.e. if we only have classical measurement error and no Berkson error then as Armstrong suggests there is an equivalence of the squared correlation between model and true and the ratio of their variances (“true” versus model). However if both classical and Berkson error are present then we can no longer assume this equivalence holds.
1.3. Also, in the scenario definitions, temporal and spatial components of lambdas and alphas are set to equal values. How does this impact on their formulas and the consequent interpretation of results?

Response: As we acknowledge in the discussion setting the temporal and spatial gammas and alphas to equal values is a limitation. However we wanted to illustrate patterns in bias, coverage probability and power and it was unclear to us how we would do this effectively by varying all 4 parameters between scenarios rather than just two. However if we want to provide performance statistics for a particular model then, provided we have a validation dataset (model versus monitor), we can estimate all 4 parameters separately and plug them into our simulation. This approach will also be useful if we wish to compare the performance of different modelling approaches or different combinations of these approaches.

In terms of the impact on formulae, within each simulation program we stored the correlation coefficients and variance ratios, both spatial and temporal, to check that on average they were close to our target values. For example for NO2 with correlation set to 0.6 and variance ratio to 0.75, on average (i.e. averaged across all 500 simulations) the temporal correlation was 0.5994, the spatial correlation was 0.5968, the temporal variance ratio was 0.7517 and the spatial variance ratio was 0.7611. Similarly reassuring statistics were obtained for all simulation programs. Even the slight upward bias noted in the spatial variance ratio (i.e.0.7611-0.75=0.0111) was never greater than 0.0240.

For complete accuracy we have changed “our main simulations nevertheless produced” to “the aim of our simulations was to produce” on Page 17 and added the following text:

“The success of incorporating these correlations and variance ratios into our “pseudo” modelled data was assessed by checks within our simulation programs. While overall these checks were reassuring they did suggest that in terms of the spatial variance ratio, the actual value introduced might be slightly higher than intended. However across all the scenarios in Tables 1 and 2 estimates of this bias (to 2 decimal places) were never more than 0.02 (e.g. spatial variance ratio 2.02 rather than 2.00).”

1.4. How much is this realistic in real-life time series data of air pollutants?

Response: As indicated above (see response to point 2 above) this is difficult to assess. Nevertheless as now mentioned (bottom of page 13 / top of page 14) the high correlation low variance ratio combination which produced non-trivial positive bias “may only occur in practice if there is a lack of independence between the Berkson component of measurement error and the modelled data. [9-10]”.

2.1 In the results, the authors find attenuation induced by classical error, with attenuation increasing for larger values of variance ratios and lower correlations between modelled and true data. This is expected and consistent. On the opposite, they find bias away from the null caused by Berkson error, with increasing bias for smaller lambda and for higher correlation. This is, to
me, extremely difficult to understand, and I was surprised that the authors didn't attempt to interpret this finding in the Discussion. They report a couple of references where bias away from the null induced by Berkson error was found (ref [8,9]). In those studies an interpretation was given by the authors, related to the log-normal distribution of the exposure, which might have violated the assumption of independence between error and modelled data. Under this situation, they argue, Berkson not only reduces power but can also induce bias, with bias increasing for higher slope between error and observed exposure data. In the present study the authors do not attempt any interpretation, and, above all, they do not explain why such non-trivial bias away from the null is highest for higher correlations between true and modelled data. Is that because, in such situations, classical error is reduced and the Berkson one dominates? If so, the authors should try to discuss it, possibly relating this to the formulas they have applied to generate simulated model data.

Response: We appreciate that for many researchers in this area the idea of positive bias arising from Berkson error unless that error is additive on a log scale (i.e. the proportional error model) is rather odd and requires further explanation. We have therefore replaced the last sentence of paragraph 3, page 13 with the following:

“However, for $\lambda=0.5$ combined with a high correlation coefficient of 0.9, bias away from the null was observed for both short and long-term exposure ranging from 27% to 44%. In trying to explain these findings we note that the scenario effectively sets the covariance between the model and “true” data equal to 1.27 times (i.e. $0.9/\sqrt{0.5}$ ) the variance of the model data. This relationship is indicative of positive bias (based on simple regression calibration)[10, 25] but may only occur in practice if there is a lack of independence between the Berkson component of measurement error and the modelled data.[9-10] While, in general Berkson error is not thought to introduce bias into the health effect estimate, some studies have shown that bias away from the null can occur due to Berkson error if additive on a log scale.[9-10].”

A lack of independence between Berkson error and modelled data violates one of the assumptions underlying what might be termed standard measurement error theory [25]. Given that our methodology allows for scenarios which violate this assumption we have therefore also changed

“…based on standard measurement error theory, [25]…” (Bottom of page 16) to “…based on a second-order regression equations,[25, 26]…”.

3.1 I would encourage the authors to interpret their findings in relation to conventional approaches used in literature to estimate short-term and long-term air pollution exposures: daily mean of central monitors, spatial kriging, LUR, spatiotemporal models, etc. It is not clear how the results of this study might be used in future epidemiological investigations to: a) improve exposure assessment, b) critically interpret epidemiological results, and c) possibly adjust "naïve" estimates to account for measurement errors.

Response: The aim of this paper was to present generalised scenarios couched in terms of familiar metrics like the correlation coefficient and variance ratio that could be readily calculated
from validation studies by modellers of air pollutants or other environmental exposures. In so doing we hope to highlight some of the dangers associated with certain metric combinations.

However as indicated in our conclusion, in terms of the future, we envisage the expanded form of our methodology (i.e. separate correlation and variance ratios, temporally and spatially) being used to compare the performance (in terms of bias, coverage probability and power) of different pollutant modelling approaches, in order to find the best model for use in a multi-level analysis of air pollution and health. This approach would not only enable comparison of some more standard models (e.g. LUR, Dispersion, satellite based etc.) but facilitate the comparison of hybrid models produced by combining outputs from different modelling approaches.

We have tried to emphasise this aspect, by slightly expanding the last sentence of the conclusion as follows:

“By allowing these factors to differ spatially and temporally, as outlined in Additional file 2, statistical simulation can be used to compare the performance (in terms of bias, coverage probability and power) of different pollutant modelling approaches (e.g. LUR, dispersion, satellite-based etc.) in order to find the best model or combination of models for use in a multi-level analysis of air pollution and health.”

Other Issues:

In reading through our manuscript we noted an error in the numbering of our last 3 references. This error has now been corrected. We also picked up a small error in our PM10 simulation programs. These programs have been re-run and the text (see highlighting), tables and figures amended accordingly. A few other minor corrections have also been made to the text (see highlighting), Additional file 2, and the footnote to tables 1 and 2 has been expanded as follows: “¶ Coefficients and standard errors are averages of their respective within-simulation estimates. §Percent bias is highlighted in bold when positive (i.e. away from the null). ‡The percentage of effect estimates that were statistically significant (p<0.05)”. We have not highlighted changes in tables or Additional files but versions with track changes can be provided if required.