

Incorporating Measurement Error from Modeled Air Pollution Exposures into Epidemiological Analyses

Evangelia Samoli¹ · Barbara K. Butland²

Published online: 5 October 2017

© Springer International Publishing AG 2017, Corrected publication October/ 2017

Abstract

Purpose of review Outdoor air pollution exposures used in epidemiological studies are commonly predicted from spatio-temporal models incorporating limited measurements, temporal factors, geographic information system variables, and/or satellite data. Measurement error in these exposure estimates leads to imprecise estimation of health effects and their standard errors. We reviewed methods for measurement error correction that have been applied in epidemiological studies that use model-derived air pollution data.

Recent findings We identified seven cohort studies and one panel study that have employed measurement error correction methods. These methods included regression calibration, risk set regression calibration, regression calibration with instrumental variables, the simulation extrapolation approach (SIMEX), and methods under the non-parametric or parameter bootstrap. Corrections resulted in small increases in the absolute magnitude of the health effect estimate and its standard error under most scenarios.

Summary Limited application of measurement error correction methods in air pollution studies may be attributed to the

absence of exposure validation data and the methodological complexity of the proposed methods. Future epidemiological studies should consider in their design phase the requirements for the measurement error correction method to be later applied, while methodological advances are needed under the multi-pollutants setting.

Keywords Air pollution · Bootstrap · Health · Measurement error · Regression calibration · SIMEX

Introduction

The health effects of air pollution have been widely investigated during the last decades under different epidemiological designs. Effects of air pollution related to short-term exposure have used time series and panel study designs, while those related to long-term exposure have been investigated through the follow up of cohorts. Until 2000, the exposure indices were typically derived from fixed monitoring stations measuring regulated pollutants in metropolitan areas, but more recently, this approach has been challenged due to the need to investigate associations in areas where the monitoring network is sparse or even absent as well as to investigate associations with non-regulated air pollutants that require study specific measuring campaigns. More importantly many researchers have highlighted the need to provide better exposure assessment in order to minimize measurement error. Although the problem of measurement error has been widely investigated in other areas of epidemiological research, such as nutritional epidemiology, it has only been addressed in environmental epidemiology relatively recently. Zeger et al. were among the first to discuss the effect of measurement error in time-series [1], followed by Brauer et al. who investigated its impact on the estimation of a threshold in the association [2].

This article is part of the Topical Collection on *Air Pollution and Health*

✉ Evangelia Samoli
esamoli@med.uoa.gr

¹ Department of Hygiene, Epidemiology and Medical Statistics, Medical School, National and Kapodistrian University of Athens, 75 Mikras Asias Str, 115 27 Athens, Greece

² Population Health Research Institute and MRC-PHE Centre for Environment and Health, St George's, University of London, London, UK

The need for improving the spatial resolution of exposure estimates has been addressed by methods using modeling techniques to estimate concentrations of pollutants in geographical grids (dispersion modeling, use of satellite data) or at points in space such as the residential address of cohort participants (through spatial regression models such as land-use regression (LUR) or universal kriging). The advent of geographic information system (GIS)-based data that are used as predictors of exposure have supported these attempts. Epidemiological research has used the estimated exposures derived from the abovementioned methods as single “true” values not accounting for the impacts of the uncertainty on the health effect estimates.

Gryparis et al. and Spirzo et al. proposed methods for measurement error correction in the presence of spatially misaligned modeled air pollution data [3, 4], while Sheppard et al. reviewed these methods [5]. Since then, measurement error correction techniques have been expanded, and new ones proposed, although only a few of these have been applied in air pollution studies.

In this paper, we briefly review the methods for measurement error correction that have been applied in air pollution epidemiological studies. The focus is not on the description of the methodological proposals to assess or correct for measurement error, but on the methods that have been applied in air pollution epidemiological research to account for the uncertainty in modeled exposure estimates in their health analyses.

Exposure Measurement Error in Air Pollution Estimation

Although true exposures to air pollution reflect personal exposures of the individual to local emission sources during different activities, the exposure assessment methods proposed so far are an approximation of the personal exposure derived from outdoor sources. This is largely due to lack of time-activity data, an ability to accurately predict or measure concentrations in different microenvironments, and the fact that exposure from outdoor sources is the most relevant for policy makers. As such, the concept of measurement error in air pollution epidemiology mainly addresses this part of the exposure. The use of modeled pollutant concentrations to estimate exposure from outdoor sources induces a combination of classical-type error linked to error in monitor measurements and in the estimation of prediction model parameters, and Berkson-type error due to the use of prediction models that have elements common across individuals (use of GIS data).

Measurement Error Issues Relevant to Air Pollution Epidemiology

There are two main types of measurement error: classical and Berkson. In short, if X is the unobserved “true” exposure

which is approximated by an observed surrogate Z , U_c is the classical error and U_b is the Berkson error, then under the additive classical error model $Z = X + U_c$ where $E[Z|X] = X$ and $\text{Var}(Z) = \text{Var}(X) + \text{Var}(U_c)$ denoting that the observed values are distributed at random around the true exposure; and under the additive Berkson error model, $X = Z + U_b$ where $E[X|Z] = Z$ and $\text{Var}(X) = \text{Var}(Z) + \text{Var}(U_b)$ denoting that the true exposures are distributed at random around the observed surrogate. Berkson error arises where Z represents an average exposure of a group with similar characteristics such as the modeled exposures for all individuals living at a single residential address. Classical and Berkson errors have different implications for effect estimates and their standard errors in health models. In the simplest case (i.e., replacing X by Z as the explanatory variable in a simple linear regression analysis), the additive measurement error leads to attenuation in the regression coefficient by $(1 - C) \times 100\%$ where $C = \text{Cov}(X, Z)/\text{Var}(Z)$. If all the measurement error is classical, then $C = \text{Var}(X)/\text{Var}(Z)$. Under this scenario, the standard error of the regression coefficient may be biased in either direction [5]. Berkson error typically leads to an unbiased estimate ($C = \text{Var}(Z)/\text{Var}(Z) = 1$) but with inflated standard errors. However the impact of either type of error depends on a number of assumptions including normality and linearity, but also the hypothesis that the error distribution does not depend on the health outcome (defined as non-differential error). The latter implies that you only measure the surrogate because the true exposure is unavailable and that the surrogate provides no additional information about the outcome (i.e., beyond that provided by the true exposure if observed).

Exposure Assessment Models and Impact of Measurement Error in Air Pollution Epidemiology Studies

A review of exposure assessment methods is presented in Hoek [6]. Briefly, modeled approaches to estimate spatially variable air pollution concentrations may be classified in three broad classes: (1) Dispersion or chemical transport models are based on atmospheric meteorology and chemistry, and model dispersion and chemical transformations in the atmosphere based on emissions data. They provide estimates of concentrations in spatial grids which may be downscaled to relatively small dimensions (recent advances provide estimates down to $100 \text{ m} \times 100 \text{ m}$ grids) taking into account atmospheric processes. Dispersion models are based on our deterministic understanding of atmospheric processes. Errors in these models primarily reflect imprecise or erroneous emissions data and inability to capture the complexities of the multiple small scale determinants either in the natural or built environment; (2) LUR and other spatial regression models are based on geographic covariates and are applied to expand in situ measurements of pollutant concentrations to large areas [7]. These

models can estimate exposure concentrations at any point in space using coordinates and thus have the ability in principle to approximate “personal” exposure better by providing estimates at subjects’ residential addresses. Both dispersion and spatial modeling approaches have been extended to incorporate temporal as well as spatial variability. Spatiotemporal LUR models that combine large numbers of correlated geographic information system (GIS) variables (by incorporating factors constructed using partially weighted least squares or other dimension reduction techniques) and smooth surfaces (using kriging approaches) have also been developed [8, 9]. As LUR models are based on measured concentrations, they reflect errors related to instrumental errors in the measurements, the number and location of monitoring sites, the limited monitoring period, and statistical uncertainty; and (3) Models incorporating satellite measurements, as Aerosol Optical Density (AOD) data are used in the estimation of particulate pollution. Spatiotemporal resolved particulate matter with aerodynamic diameter less than $2.5\ \mu\text{m}$ ($\text{PM}_{2.5}$) exposures by AOD calibration has been developed [10]. Satellite data have the advantage of being widely available, and the errors reflect the estimation of concentrations at heights relevant for humans, the lack of spatial precision, and interference by cloud cover, nearby water, etc.

Combining various approaches to exposure assessment is believed to lead to optimizing their advantages and minimizing their respective disadvantages. More recently hybrid models are being developed that use combinations of the main modeling exposure methods outlined above. Dispersion model outputs have been used as input for LUR [11], while De Hoogh et al. have developed LUR models incorporating both estimates from dispersion models and AOD or NO_2 satellite measurements to estimate $\text{PM}_{2.5}$ and NO_2 concentrations at $100\ \text{m} \times 100\ \text{m}$ spatial resolution [12]. Furthermore, machine learning methods have been applied to incorporate a large number of predictors for optimal estimation of pollution concentrations. Reid et al. estimated daily exposure to $\text{PM}_{2.5}$ through data-adaptive machine learning methods [13], while LUR and remote sensing data have been combined in the USA using neural network methods [14].

Several simulation studies have addressed the amount of bias introduced by different approaches to specific design types: Kim et al. compared the use of nearest monitor data to kriging estimates in a cohort study to show that the bias in the health effect estimates are dependent on the amount of spatial dependence in the underlying exposure distribution [15]; Szpiro et al. used LUR predictions in a linear disease model to conclude that more accurate exposure prediction at subject’s location does not necessarily lead to improved health effect estimation [16]; Basagaña et al. reported in a longitudinal study that the bias in the health estimate from using LUR predictions depends on the number of air pollution measurement sites, the number of available predictors for model

selection, and the amount of explainable variability in the true exposure [17]; and finally, Butland et al. evaluated dispersion estimates vs monitor measurements in time-series concluding that model data may lead to appreciable bias in health effect estimates [18], while Dionisio et al. investigated the effect of measurement error in Poisson models including multiple pollutants warning that attenuation may be substantial [19, 20]. Comparison of health effects derived from different exposure estimates have also been reported as in Sellier et al. who used data from a cohort study to show that the use of monitor data, dispersion estimates, or LUR estimates may lead to different health effect estimates when a heterogeneous pollutant such as NO_2 is under investigation [21].

Measurement Error Correction in Air Pollution Epidemiological Studies

We reviewed *Pubmed* from 2011 onwards to February 2017 using as keywords “measurement error,” “air pollution,” and “health” to identify methods that have been applied in the last years and present the associated results. From the 51 initially identified papers, there were few methodological papers proposing methods for correcting the health estimates when using modeled air pollution exposures, and even fewer epidemiological studies have incorporated these methods in their publications. As modeled exposures are more widely applied in prospective study designs, all measurement error correction applications but one were in cohort studies [22•, 23•, 24•, 25•, 26•, 27•, 28•]. We included one air pollution panel study with measurement error correction, though slightly outside the remit of our review (pollution data not model-based), because the authors apply a method not presented in any of the other identified studies [29•]. We briefly present below the methods applied in the identified papers and their results.

Regression Calibration and Extensions

Regression calibration (RC) is a well-established method of correcting for measurement error in epidemiological regression models [30]. In general, for a subset of study subjects, it requires information on both the surrogate and a more accurate measure of exposure. It is the linear regression of the latter on the former which facilitates the correction of health effect estimates and their 95% confidence intervals. This correction can be applied post-analysis by calculating and correcting for the attenuation factor C or within the main analysis by re-calibration of the surrogate variable [31]. In the Netherlands Cohort Study on Diet and Cancer, Hart et al. used a traditional RC approach, informed by data from a validation study comparing personal and ambient measures of traffic related air pollution, to estimate measurement-error-adjusted hazard ratios for lung cancer and long-term exposure to each of $\text{PM}_{2.5}$, black smoke (BS), and

NO_2 [22•]. As noted by the authors, adjustment for measurement error, moved the hazard ratios (HR) for all incident lung cancer and each of four subtypes (squamous, small-cell, large-cell, adenocarcinoma) further from the null, with increases of 0.0–3.3% for BS and 9.7–37.2% for $\text{PM}_{2.5}$ (Table 1) [22•]. The magnitude of the percent increase in the width of the confidence intervals was larger, with increases of 10.2–23.3% for BS and 108.0–216.8% for $\text{PM}_{2.5}$ [22•].

Risk set regression calibration (RRC) is an extension of RC for use in survival analysis with time-varying exposures which allows the re-calibration of the surrogate variable to be updated over time and with changes in the risk set (i.e., those subjects at risk). A sandwich variance estimator is used to calculate Wald-type asymptotic confidence intervals and p -values. First described by Liao et al. [32], this method of measurement error correction was applied by Hart et al. [23•], when investigating the long-term effects of modeled $\text{PM}_{2.5}$ exposure on all-cause mortality in the Nurses' Health Study. The authors noted that measurement error correction increased the hazard ratio by 4.4% and doubled the width of the 95% confidence interval (Table 1) [23•].

Another extension of RC involves the long-standing technique for measurement error correction of instrumental variables. A variable is considered a good “instrument” if it is correlated with the true exposure, independent of the error term in the surrogate, and provides no more information about the outcome than the true exposure. Regression calibration with instrumental variables (RCIV) has been used to provide methods of error correction which can be applied in epidemiological models of increasing complexity. One approach involves a post-analysis correction based on the assumption that both the true and surrogate variables can be predicted using identical linear combinations of the instrumental variables (RCIV1) [29•]. In RCIV1, correction is based on parameter estimates obtained from regressing each surrogate on its corresponding instrumental variable or variables. Stand et al. also proposed an extension of RCIV1 to facilitate the post-analysis correction of a two-pollutant model with pollutant interaction terms plus interactions with error-free covariates. In that approach, they obtained parameter estimates using multi-level models with random intercepts for each subject and assuming a spatial power structure for within-subject error [29•]. They illustrated their method using data from a panel study of asthmatic children investigating the relationship between LTE4, an inflammatory biomarker, and the short-term exposure to particulates and second-hand smoke (SHS), and their interaction. They used fixed-site outdoor $\text{PM}_{2.5}$ and urinary cotinine as observed instrumental variables in place of the unobserved exposures of interest i.e., personal exposure to $\text{PM}_{2.5}$ and SHS. The surrogate variables were error-prone measurements of $\text{PM}_{2.5}$ and SHS, based on personal monitoring, which were only available for a limited number of days. Change in effect estimates of interest before and after correction were not reported.

Simulation Extrapolation (SIMEX)

If the classical measurement error variance ($\sigma_e^2 = \text{Var}(U_c)$) is known or can be estimated based on an external validation dataset, then SIMEX works by adding classical errors of increasing magnitude of the form $\lambda\sigma_e^2$ ($\lambda = 0, \dots, 2$) to the surrogate exposure (Z) and calculating health effect estimates. A regression of mean health effect versus error is then used to identify the health effect estimate corresponding to zero measurement error (i.e., $\lambda = -1$) [33]. Recently, Alexeeff et al. proposed an extension of this technique, which they refer to as spatial SIMEX, to take account of the fact that classical errors may be spatially correlated [24•]. To demonstrate this technique, they used data from a Massachusetts birth cohort and investigated the association between birthweight and $\text{PM}_{2.5}$ exposure in different trimesters of pregnancy. However, they first needed to specify the percentage of spatial measurement error that was classical and this value was not identifiable. Assuming a value of 80% (the value at which spatial SIMEX performed best in their simulation results), the authors noted that post correction, the negative relationship between $\text{PM}_{2.5}$ and birthweight in the second and third trimesters increased in absolute value by 57% and 41%, respectively, but the width of 95% confidence intervals decreased markedly by 90% in both cases [24•].

Parameter and Non-parametric Bootstrap

Szpiro et al. noted that the parameter estimation and spatial smoothing inherent in the statistical modeling of pollution data (e.g., LUR) introduce classical-like and Berkson-like error, respectively [4]. To correct for such error, bootstrapping methods have been used [4, 25•, 26•, 27•, 28•]. In its simplest form, bootstrapping builds up the sampling distribution of a sample estimate (e.g., the hazard ratio) by taking multiple random samples with replacement from the observed dataset [34]. The non-parametric bootstrap differs from the parameter bootstrap in that the former assumes the pollution exposure surface is fixed and aims to accurately reflect variation due to the sampling of monitor-site locations and study subjects whereas the latter assumes that monitor site locations and the study sample are fixed and focuses instead on accurately reflecting random variation in the pollution exposure surface [4, 28•].

Bergen et al. used the parameter bootstrap to correct for measurement error when investigating the effects of long-term exposure to various modeled $\text{PM}_{2.5}$ components on carotid-intima-media thickness within the Multi-Ethnic Study of Atherosclerosis (MESA) [25•]. While there was no obvious change in effect estimates, the width of confidence intervals increased by 56% for silicon and 48% for sulfur after accounting for errors in the exposure models (Table 1). Results from using a simplified version of the parameter

Table 1 Air pollution epidemiological studies that have applied measurement error correction methods for the effect estimate

References ordered by correction method	Health outcome	Exposure assessment	Uncorrected effect estimate (95% CI)	ME correction method	Corrected effect estimate (95% CI)
Hart et al. 2015 [22•] Netherlands Cohort Study on Diet and Cancer	Lung cancer	Long-term average exposure to black smoke (BS) and PM _{2.5} from LUR models	Hazard ratio (HR) = 1.16 (1.02, 1.32) per 10 µg/m ³ BS HR = 1.17 (0.93, 1.47) per 10 µg/m ³ PM _{2.5}	Regression calibration	HR = 1.19 (1.02, 1.39) per 10 µg/m ³ BS HR = 1.37 (0.86, 2.17) per 10 µg/m ³ PM _{2.5}
Hart et al. 2015 [23•] Nurses' Health Study, USA	All-cause mortality	12-month average exposure to PM _{2.5} from spatiotemporal models	HR = 1.13 (1.05, 1.22) per 10 µg/m ³ PM _{2.5}	Risk set calibration	HR = 1.18 (1.02, 1.36) per 10 µg/m ³ PM _{2.5}
Alexeeff et al. 2016 [24•] Birth cohort, Massachusetts, USA	Birthweight	PM _{2.5} exposure during the second and third gestational trimesters using a universal kriging model with Matérn residuals	- 5.04 g (- 8.02, - 2.05) per 1 µg/m ³ second trimester PM _{2.5} ; - 3.49 g (- 6.08, - 0.89) per 1 µg/m ³ third trimester PM _{2.5}	Spatial SIMEX	- 7.90 g (- 8.20, - 7.61) per 1 µg/m ³ second trimester PM _{2.5} ; - 4.91 g (- 5.17, - 4.66) per 1 µg/m ³ third trimester PM _{2.5}
Bergen et al. 2013 [25•] Multi-Ethnic Study of Atherosclerosis, USA	Carotid intima-media thickness	Annual average PM _{2.5} component concentrations using universal kriging with partial least squares to select geographic covariates	0.001 (- 0.03, 0.03) mm per 1 µg/m ³ PM _{2.5} elemental carbon (EC) content 0.025 (0.01, 0.04) mm per 1 µg/m ³ PM _{2.5} organic carbon (OC) content 0.408 (0.25, 0.57) mm per 1 ng/m ³ PM _{2.5} silicon (Si) content 0.055 (0.022, 0.088) mm per 1 µg/m ³ PM _{2.5} sulfur (S) content	Parameter bootstrap Partial parametric bootstrap	Correcting for classical-like and Berkson-like error: 0.001 (- 0.03, 0.03) mm per 1 µg/m ³ EC 0.025 (0.01, 0.04) mm per 1 µg/m ³ OC 0.408 (0.16, 0.66) mm per 1 ng/m ³ Si 0.055 (0.006, 0.104) mm per 1 µg/m ³ S Correcting for Berkson-like error only: 0.001 (- 0.03, 0.03) mm per 1 µg/m ³ EC 0.025 (0.01, 0.04) mm per 1 µg/m ³ OC 0.408 (0.16, 0.66) mm per 1 ng/m ³ Si 0.055 (0.006, 0.104) mm per 1 µg/m ³ S
Szpiro & Paciorek 2013 [26•] Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air), USA	Elevated left ventricular mass index	LUR estimates of residential concentrations of NOx	0.66 (- 0.56, 1.88) g/m ² per 10 ppb NOx	Non-parametric bootstrap (primarily Berkson-like error correction) Bias correction (classical-like error correction) and non-parametric bootstrap (primarily Berkson-like error correction)	0.66 (- 0.63, 1.95) g/m ² per 10 ppb NOx 0.68 (- 0.65, 2.01) g/m ² per 10 ppb NOx

Table 1 (continued)

References ordered by correction method	Health outcome	Exposure assessment	Uncorrected effect estimate (95% CI)	ME correction method	Corrected effect estimate (95% CI)
Bergen et al. 2016 [27•] Sister Study of the National Institute of Environmental Health Sciences, USA	Systolic blood pressure	NO ₂ and PM _{2.5} exposures from low-rank kriging penalized regression models	Two pollutant model results: 3.55 (2.06, 5.04) mmHg per 10 µg/m ³ PM _{2.5} - 1.12 (- 1.60, - 0.64) mmHg per 10 ppb NO ₂	Bias correction and non-parametric bootstrap	Two pollutant model results: 4.96 (1.76, 8.16) mmHg per 10 µg/m ³ PM _{2.5} - 1.71 (- 2.91, - 0.51) mmHg per 10 ppb NO ₂
Keller et al. 2017 [28•] Birth cohort, Georgia, USA	Birthweight	PM _{2.5} exposure during the third gestational trimester using a spatiotemporal model	- 2.4 g (- 3.9, - 0.8) per 1 µg/m ³ third trimester PM _{2.5}	Non-parametric bootstrap Parameter bootstrap	- 2.5 g (- 4.2, - 0.8) per 1 µg/m ³ PM _{2.5} - 2.4 g (- 4.2, - 0.6) per 1 µg/m ³ PM _{2.5}

LUR Land Use Regression

bootstrap (i.e., the “partial parametric bootstrap”) which only adjusts for Berkson-like error, were identical (Table 1).

In a related study, Szpiro & Paciorek proposed both a one- and two-stage process involving bootstrapping [26•]. In the one-stage process, a non-parametric bootstrap was used to obtain both bias-corrected effect estimates and adjusted standard errors. In the two-stage process, an asymptotic correction factor was used to adjust for bias (caused by the classical-like component) followed by a non-parametric bootstrap to adjust standard errors for both error types. To illustrate their methods, they used data from MESA and examined the relationship between elevated left ventricular mass and long-term exposure to NO_x. The health regression coefficient, though unaffected by measurement error correction under the one-stage process, increased by 3% under the two-stage process (Table 1). Both methods resulted in a widening of the 95% confidence interval by 6% and 9%, respectively [26•].

Later, Bergen et al. demonstrated that for exposure data predicted using penalized splines, this two-stage process can be extended to correct for measurement error in a multi-pollutant health effects model [27•]. With data from the Sister Study of the National Institute of Environmental Health Sciences, they investigated the mutually adjusted long-term effects of exposure to NO₂ and PM_{2.5} on systolic blood pressure. Following measurement error correction, both regression coefficients in the main analysis moved further from the null (the coefficient for NO₂ became more negative and the coefficient for PM_{2.5} more positive) and 95% confidence intervals widened substantially (Table 1, personal communication). A secondary analysis restricted to the north east region, however, showed little effect of measurement error corrections with, if anything, hazard ratios moving toward rather than away from 1 [27•].

While the above methods focused on spatial measurement error [4, 25•, 26•, 27•], Keller et al. [28•] used extensions of both the non-parametric and parameter bootstrap to adjust for spatiotemporal measurement error in a study of birthweight and gestational trimester-specific PM_{2.5}. They found that the regression coefficient linking birthweight and PM_{2.5} exposure in the third trimester was unchanged by the parameter bootstrap but increased by 5.9% under the non-parameter bootstrap, although under both methods confidence intervals widened by 16% and 10%, respectively (Table 1).

Discussion

Although measurement error correction approaches in epidemiological models have been proposed for some time, they have only recently received attention in their application in air pollution studies. We reviewed recent relevant literature for application of these methods and identified only eight publications that have applied a variety of methods. All but one

identified studies were cohorts, six of which were from the USA. The measurement error correction methods applied included regression calibration, risk set regression calibration, regression calibration with instrumental variables, SIMEX, the parameter bootstrap, and the non-parametric bootstrap. While our focus here has been on cohort studies, based on what has been worked on in the literature, it is worth noting that measurement error may be more of an issue in terms of health effect estimation in time-series or panel studies, where the temporal component of additive classical measurement error is not reduced by averaging over time [18].

Findings both before and after measurement correction from recent studies (2011 and later), when provided in the relevant papers, are summarized in Table 1. Overall measurement error correction tended to increase the absolute magnitude of the health effect estimate, illustrating the tendency for additive classical error to bias effect estimates toward the null. However, this pattern was not so obvious in those studies using the non-parametric or parameter bootstrap. Bergen et al. [25•] attributed this to the absence of classical-like error in their exposure data as was shown by a comparison of effect estimates corrected for both error types with those corrected for Berkson-like error alone (Table 1). Similarly, while Szpiro & Paciorek noted little difference between the one- and two-stage process in their data analysis, simulations provided evidence that their two-stage approach improved estimation and inference [26•].

The effects of measurement error adjustment on standard errors was more difficult to predict and ranged from the narrowing of confidence intervals seen in the Alexeeff et al. study under spatial SIMEX [24•] to the widening of confidence intervals observed under RC, RRC, and bootstrapping [22•, 23•, 25•, 26•, 27•, 28•]. There are various possible explanations for these differences: classical error can bias standard errors in either direction [5], spatial correlation in Berkson-like error can lead to underestimated standard errors [4, 25•], and measurement error correction itself may introduce additional variability [22•]. However, it is difficult to draw conclusions from such comparisons as we are not comparing like with like, due to underlying framework differences between methods.

Each correction method considered in this review has strong assumptions that result in different advantages and disadvantages. Spatial SIMEX is dependent upon the underlying distribution of the true exposure surface and applied using different scenarios as to the proportion of total measurement error that is either classical or Berkson-like and depending on these assumptions provides a range of corrected effect estimates. The regression calibration approach has strong assumptions on the normality and linearity of the true and the “measured” part of exposure that in principle may be overcome by appropriate transformations, but in addition, it assumes that the measurement error variance is small. Although extensions to overcome these restrictions have been proposed (such as semiparametric efficient estimator), they

have not been applied in air pollution epidemiological settings. Finally, bootstrap methods rely on the available exposure surface, assuming representativeness of the underlying true surface.

Our review findings illustrate that in a single-pollutant epidemiological model, the presence of additive classical error typically biases the health effect estimate toward the null. However, in a multi-pollutant model, the direction of bias due to multi-pollutant measurement error is far less predictable especially when the pollutants are highly correlated [35]. The biases themselves may also be more severe [27•]. In addition to the non-parametric bootstrap [27•], SIMEX can be extended to provide measurement error adjustment in certain multi-pollutant settings [24•]. Further novel methodological research is needed that would also provide epidemiologists with publically available computer codes for ease of implementation. Such extensions become important since as pointed out by Dominici et al. [36], the focus of health effect estimation is moving away from a single-pollutant approach, and recent epidemiological findings related to correlated particles’ physical and chemical characteristics need to be correctly accounted for.

With the exception of the bootstrap approaches, the measurement error correction methods presented in this paper require additional data in the form of a validation study or dataset, and if that study/dataset is too small, it may compromise the adequacy of adjustment. In the case of SIMEX, external data are required in order to estimate the total measurement error variance in the surrogate variable. To overcome this issue, the held-out monitors approach has been suggested, under which a spatial model is fit to the difference between predicted and measured values at the leave-out monitors. The need for such information is therefore something that researchers should consider when designing an epidemiological study of air pollution and health. Finally, although computer libraries for regression calibration and SIMEX are incorporated in standard statistical software, their underlying assumptions and complexity may have hindered their application. On the other hand, the bootstrap methods add a degree of complexity as they also require the simulation of the exposure surface and not just the correction in the epidemiological analysis stage. Air pollution epidemiological research should move forward by reporting effect estimates under a variety of correction methods depending on the focus of the study, its design and sample size (as for example, the size of administrative cohorts prohibits the application of bootstrap methods), the availability of validation data, and the nature of the exposure distribution.

Conclusions

In air pollution epidemiology, effect estimates tend to be small but have important implications for population health, due to

the fact that we are all exposed throughout our lifetimes. Therefore, accurate estimation and the elimination of bias due to measurement error is particularly important. Yet, personal air pollution exposures are very challenging to quantify accurately, and few studies currently correct for measurement error. Barriers for the adoption of these methods in the mainstream literature may only get worse as measurement error correction methods of increasing complexity may be required as the focus moves to the multi-pollutant setting, and there is a need to separate the effects of highly correlated pollutants.

Acknowledgements The authors would like to thank Dr. Richard Atkinson for his valuable review and comments on this paper.

Compliance with Ethical Standards

Conflict of Interest Evangelia Samoli and Barbara Butland declare that they have no conflict of interest.

Human and Animal Rights and Informed Consent This article does not contain any studies with human or animal subjects performed by any of the authors.

References

Papers of particular interest, published recently, have been highlighted as:

- Of importance
1. Zeger SL, Thomas D, Dominici F, Samet JM, Schwartz J, Dockery D, et al. Exposure measurement error in time-series studies of air pollution: concepts and consequences. *Environ Health Perspect*. 2000;108:419–26.
 2. Brauer M, Brumm J, Vedal S, Petkau AJ. Exposure misclassification and threshold concentrations in time series analyses of air pollution health effects. *Risk Anal*. 2002;22:1183–93.
 3. Gryparis A, Paciorek CJ, Zeka A, Schwartz J, Coull BA. Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics*. 2009;10:258–74.
 4. Szpiro AA, Sheppard L, Lumley T. Efficient measurement error correction with spatially misaligned data. *Biostatistics*. 2011;12:610–23.
 5. Sheppard L, Burnett RT, Szpiro AA, Kim S-Y, Jerrett M, Pope CA 3rd, et al. Confounding and exposure measurement error in air pollution epidemiology. *Air Qual Atmos Health*. 2012;5:203–16.
 6. Hoek G: Methods for assessing long-term exposures to outdoor air pollutants. *Curr Environ Health Rep*. 2017. in press.
 7. Eeftens M, Beelen R, de Hoogh K, Bellander T, Cesaroni G, Cirach M, et al. Development of land use regression models for PM(2.5), PM(2.5) absorbance, PM(10) and PM(coarse) in 20 European study areas; results of the ESCAPE project. *Environ Sci Technol*. 2012;46:11195–205.
 8. Sampson PD, Richards M, Szpiro AA, Bergen S, Sheppard L, Larson TV, et al. A regionalized national universal kriging model using partial least squares regression for estimating annual PM2.5 concentrations in epidemiology. *Atmos Environ*. 2013;75:383–92.
 9. Keller JP, Olives C, Kim SY, Sheppard L, Sampson PD, Szpiro AA, et al. A unified spatiotemporal modeling approach for predicting concentrations of multiple air pollutants in the multi-ethnic study of atherosclerosis and air pollution. *Environ Health Perspect*. 2015;123:301–9.
 10. Kloog I, Nordio F, Coull BA, Schwartz J. Incorporating local land use regression and satellite aerosol optical depth in a hybrid model of spatiotemporal PM2.5 exposures in the Mid-Atlantic states. *Environ Sci Technol*. 2012;46:11913–21.
 11. Akita Y, et al. Large scale air pollution estimation method combining LUR and chemical transport modeling. *Environ Sci Technol*. 2014;48:4452.
 12. de Hoogh K, Gulliver J, Donkelaar AV, Martin RV, Marshall JD, Bechle MJ, et al. Development of West-European PM2.5 and NO2 land use regression models incorporating satellite-derived and chemical transport modelling data. *Environ Res*. 2016;151:1–10.
 13. Reid CE, Jerrett M, Tager IB, Petersen ML, Mann JK, Balmes JR. Differential respiratory health effects from the 2008 northern California wildfires: a spatiotemporal approach. *Environ Res*. 2016;150:227–35.
 14. Di Q, Kloog I, Koutrakis P, Lyapustin A, Wang Y, Schwartz J. Assessing PM2.5 exposures with high spatiotemporal resolution across the continental United States. *Environ Sci Technol*. 2016;50:4712–21.
 15. Kim S-Y, Sheppard L, Kim H. Health effects of long-term air pollution: influence of exposure prediction methods. *Epidemiology*. 2009;20:442–50.
 16. Szpiro AA, Paciorek CJ, Sheppard L. Does more accurate exposure prediction necessarily improve health effect estimates? *Epidemiology*. 2011;22:680–5.
 17. Basagaña X, Aguilera I, Rivera M, Agis D, Foraster M, Marrugat J, et al. Measurement error in epidemiologic studies of air pollution based on land-use regression models. *Am J Epidemiol*. 2013;178:1342–6.
 18. Butland BK, Armstrong B, Atkinson RW, Wilkinson P, Heal MR, Doherty RM, et al. Measurement error in time-series analysis: a simulation study comparing modelled and monitored data. *BMC Med Res Methodol*. 2013;13:136.
 19. Dionisio KL, Chang HH, Baxter LK. A simulation study to quantify the impacts of exposure measurement error on air pollution health risk estimates in copollutant time-series models. *Environ Health*. 2016;15:114.
 20. Dionisio KL, Baxter LK, Chang HH. An empirical assessment of exposure measurement error and effect attenuation in bipollutant epidemiologic models. *Environ Health Perspect*. 2014;122:1216–24.
 21. Sellier Y, Galineau J, Hulin A, Caini F, Marquis N, Navel V, et al. EDEN mother–child cohort study group. Health effects of ambient air pollution: do different methods for estimating exposure lead to different results? *Environ Int*. 2014;66:165–73.
 22. Hart JE, Spiegelman D, Beelen R, Hoek G, Brunekreef B, Schouten LJ, et al. Long-term ambient residential traffic-related exposures and measurement error-adjusted risk of incident lung cancer in the Netherlands Cohort Study on Diet and Cancer. *Environ Health Perspect*. 2015;123:860–6. **This study includes an application of regression calibration.**
 23. Hart JE, Liao X, Hong B, Puett RC, Yanosky JD, Suh H, et al. The association of long-term exposure to PM2.5 on all-cause mortality in the Nurses' Health Study and the impact of measurement-error correction. *Environ Health*. 2015;14:38. **This study includes an application of risk set regression calibration.**
 24. Alexeeff SE, Carroll RJ, Coull B. Spatial measurement error and correction by spatial SIMEX in linear regression models when using predicted air pollution exposures. *Biostatistics*. 2016;17:377–89. **This study includes an application of spatial SIMEX.**
 25. Bergen S, Sheppard L, Sampson PD, Kim S-Y, Richards M, Vedal S, et al. A national prediction model for PM2.5 component exposures and measurement error-corrected health effect inference. *Environ Health Perspect*. 2013;121:1017–25. **This study includes an application of both the parameter bootstrap and the partial parametric bootstrap.**

26. Szpiro AA, Paciorek CJ. Measurement error in two-stage analyses, with application to air pollution epidemiology. *Environmetrics*. 2013;24:501–17. **This study includes an application of the non-parametric bootstrap.**
27. Bergen S, Sheppard L, Kaufman JD, Szpiro AA. Multipollutant measurement error in air pollution epidemiology studies arising from predicting exposures with penalized regression splines. *Appl Stat*. 2016;65:731–53. **This study includes an application of the non-parametric bootstrap in two-pollutant models.**
28. Keller JP, Chang HH, Strickland MJ, Szpiro AA. Measurement error correction for predicted spatiotemporal air pollution exposures. *Epidemiology*. 2017;28:338–45. **This study includes an application of both the non-parametric bootstrap and the parameter bootstrap.**
29. Strand M, Sillau S, Grunwald GK, Rabinovitch N. Regression calibration with instrumental variables for longitudinal models with interaction terms, and application to air pollution studies. *Environmetrics*. 2015;26:393–405. **This study includes an application of regression calibration using instrumental variables.**
30. Spiegelman D, McDermott A, Rosner B. Regression calibration method for correcting measurement-error bias in nutritional epidemiology. *Am J Clin Nutr*. 1997;65:1179S–86S.
31. Bateson TF, Wright JM. Regression calibration for classical exposure measurement error in environmental epidemiology studies using multiple local surrogate exposures. *Am J Epidemiol*. 2010;172:344–52.
32. Liao X, Zucker DM, Li Y, Spiegelman D. Survival analysis with error-prone time-varying covariates: a risk set calibration approach. *Biometrics*. 2011;67:50–8.
33. Stefanski LA, Cook J. Simulation extrapolation: the measurement error jackknife. *J Am Stat Assoc*. 1995;90:1247–56.
34. Kosmidis I. Bias in parametric estimation: reduction and useful side-effects. *WIREs Comput Stat*. 2014;6:185–96.
35. Fung KY, Krewski D. On measurement error adjustment methods in Poisson regression. *Environmetrics*. 1999;10:213–24.
36. Dominici F, Peng RD, Barr CD, Bell ML. Protecting human health from air pollution: shifting from a single-pollutant to a multi-pollutant approach. *Epidemiology*. 2010;21:187–94.