RESEARCH ARTICLE

# Accounting for measurement error to assess the effect of air pollution on omic signals

**Erica Ponzi**[1,2,3]*, **Paolo Vineis**[3,4], **Kian Fan Chung**[5,6], **Marta Blangiardo**[3]

**1** Department of Biostatistics, Epidemiology, Biostatistics and Prevention Institute, University of Zürich, Hirschengraben 84, 8001 Zürich, Switzerland, **2** Department of Biostatistics, Oslo Center for Epidemiology and Biostatistics, University of Oslo, Norway, **3** Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, United Kingdom, **4** Italian Institute for Genomic Medicine (IIGM), Turin, Italy, **5** National Heart and Lung Institute, Imperial College London, United Kingdom, **6** Royal Brompton and Harefield NHS Trust, London, United Kingdom

* erica.ponzi@uzh.ch

## Abstract

Studies on the effects of air pollution and more generally environmental exposures on health require measurements of pollutants, which are affected by measurement error. This is a cause of bias in the estimation of parameters relevant to the study and can lead to inaccurate conclusions when evaluating associations among pollutants, disease risk and biomarkers. Although the presence of measurement error in such studies has been recognized as a potential problem, it is rarely considered in applications and practical solutions are still lacking. In this work, we formulate Bayesian measurement error models and apply them to study the link between air pollution and omic signals. The data we use stem from the "Oxford Street II Study", a randomized crossover trial in which 60 volunteers walked for two hours in a traffic-free area (Hyde Park) and in a busy shopping street (Oxford Street) of London. Metabolomic measurements were made in each individual as well as air pollution measurements, in order to investigate the association between short-term exposure to traffic related air pollution and perturbation of metabolic pathways. We implemented error-corrected models in a classical framework and used the flexibility of Bayesian hierarchical models to account for dependencies among omic signals, as well as among different pollutants. Models were implemented using traditional Markov Chain Monte Carlo (MCMC) simulative methods as well as integrated Laplace approximation. The inclusion of a classical measurement error term resulted in variable estimates of the association between omic signals and traffic related air pollution measurements, where the direction of the bias was not predictable a priori. The models were successful in including and accounting for different correlation structures, both among omic signals and among different pollutant exposures. In general, more associations were identified when the correlation among omics and among pollutants were modeled, and their number increased when a measurement error term was additionally included in the multivariate models (particularly for the associations between metabolomics and $NO_2$).

OPEN ACCESS

**Citation:** Ponzi E, Vineis P, Chung KF, Blangiardo M (2020) Accounting for measurement error to assess the effect of air pollution on omic signals. PLoS ONE 15(1): e0226102. https://doi.org/10.1371/journal.pone.0226102

**Editor:** Jaymie Meliker, Stony Brook University, Graduate Program in Public Health, UNITED STATES

**Received:** September 6, 2019

**Accepted:** November 19, 2019

**Published:** January 2, 2020

**Data Availability Statement:** The data underlying the results presented in the study, as well as the minimal dataset to reproduce the analysis on the paper, are available on request from the International Agency for Research on Cancer (IARC) and their use and availability is regulated by the Exposomics Steering Board and the IARC Ethical Committee. Interested researchers can request access to the data by contacting Pietro Ferrari from the IARC at FerrariP@iarc.fr. Findings from the study can be reproduced in their entirety by obtaining the data and following the protocol reported in the Methods section, or the scripts provided as supplementary material. We confirm that the authors did not have any special access privileges to this data that others would not have.

**Funding:** This work was supported by the grant FP7 of the European Commission 'Enhanced exposure assessment and omic profiling for high priority environmental exposures in Europe' (EXPOsOMICS grant 308610 to PV). Erica Ponzi was funded by a Doc.Mobility grant from the Swiss National Science Foundation (SNSF grant P1ZHP2_178207).

**Competing interests:** The authors have declared that no competing interests exist.

## 1 Introduction

Health effects of air pollution are a major public health issue and have received increasing attention over the past decades [1, 2, 3]. In this context, the reliable estimation of risk factors and associations between environmental exposures and health conditions requires the collection of a large amount of exposure data on a relatively high number of study subjects, which is often impractical and subject to several sources of error or imprecision. This can lead not only to the presence of bias in the estimation of parameters relevant to the study but also to inaccurate conclusions when evaluating associations among pollutants, disease risk and biomarkers. Although the presence of measurement error in such studies has been discussed in the recent literature and is now recognized as a potential problem [4, 5], it is often not accounted for in standard analyses, as pointed out in [6, 7].

Different approaches to the problem have been adopted and different methods and techniques are available in the literature, for instance [8] suggested a semi-parametric approach to different types of error in radiation data, [9] proposed a Bayesian hierarchical model to retrieve error-free estimates of the health effect, [10] worked on a quantification of measurement error effect via validation studies, and [11] proposed a method to account for the error in time-series studies on air pollution. Studies focusing on the effect of traffic-related air pollution (TRAP) are particularly challenging on this matter and often rely on surrogate measures of pollutants, as well as on the approximation of personal exposures. [12] identified three main sources of error in the TRAP exposure assessment, namely a) the difference between measured and true ambient exposure levels, b) the difference between aggregate personal exposure and the exposure of a given individual, which are mostly due to approximation and classified as Berkson errors, and c) the difference between ambient concentration and average personal exposure, classified as a classical error. In this context, different methods have been proposed and a wide set of measurement error techniques have been employed, including among others the simulation extrapolation algorithm [13], regression calibration and validation data [14] and Bayesian methods [9, 15]. On the other hand, none of these studies was applied to omics data, as they focused on estimating disease risks and did not include any molecular data.

In the present study, we propose to apply measurement error techniques to correct for error in environmental exposures when considering their association with high-throughput molecular data. This is particularly challenging due to the high dimensionality of the data, as well as to the correlation among omics sampled from the same individual. We use a Bayesian framework to address the problem, which provides a very flexible way to account for measurement error and model different error types and dependency structures in the data. In particular, Bayesian hierarchical models seem ideal in these context, as they provide a straightforward way to include dependency between exposures, but also between different response variables. Moreover, the possibility to include prior knowledge on the error components can result in better models and more accurate estimations. Additionally, the possibility of modelling several fixed and random effects, as well as different link functions, adds flexibility and general applicability to the methods.

In this paper, we apply this approach to the Oxford Street II study, a randomized crossover trial where omics and air pollution measurements are employed to investigate the association between short-term exposure to traffic related air pollution and perturbation of different omic signals [16, 17]. We implement error-corrected models in a classical measurement error framework and generalize such models to account for dependencies among pollutants, as well as among response omic variables. This provides a novel way of dealing with high-dimension omic data, by including them into a Bayesian hierarchical formulation. The possibility to model more omic signals at the same time also allows to account for dependency among signals. Moreover, the inclusion of a measurement error term, which is straightforward and flexible thanks to the hierarchical formulation, has not been proposed so far in the presence of high-throughput biological data.

We implement our models using Monte Carlo Markov Chain (MCMC) in `JAGS`, but to increase the speed of the computation, we also use the integrated nested Laplace approximation approach (INLA) [18], which has recently been used to implement measurement error models, for example in [19] and [20].

The remainder of this paper is structured as follows: we first describe the study and the model to assess the association between different air pollutants and omic measurements, focusing on metabolic pathways. The paper then illustrates the Bayesian hierarchical model we formulate to account for measurement error by including a classical error (see Section 3 for definition and theoretical consideration on classical measurement error). We expand such model to a multi-response model, accounting for a dependency structure among different omic signals, and to a multi-variate model to account for dependency among different pollutants. We then show the results based on the data set from the Oxford Street II study and finally conclude with several discussion points and potential expansion of the proposed method.

## 2 Metabolic pathways in the Oxford Street II study

### 2.1 The study

The data we use here stem from the Oxford Street II Study, a randomized crossover trial within the EXPOsOMICS consortium [21]. In this study, 60 volunteers walked for two hours in a traffic-free area (Hyde Park) and in a busy shopping street (Oxford Street) of London. The walking experiments were performed on non-rainy weekdays only, from November to March, to avoid confounding from rain or pollen. Participants were divided into three groups: 1) healthy volunteers (n = 20) with a normal lung function and without a history of ischemic heart disease (IHD); 2) subjects with chronic obstructive pulmonary disease (COPD) (n = 20), without a history of IHD; and 3) subjects with clinically stable IHD over the past six months (n = 20) without COPD. Healthy participants were recruited using advertising in public areas within the Royal Brompton Hospital. Individuals with COPD or IHD were recruited from existing databases or outpatient respiratory and cardiology clinics at the Royal Brompton and Harefield NHS Foundation Trust. All current smokers or former smoker for less than 12 months were excluded, as well as people with high occupational levels of TRAP. Inclusion criteria based on age and forced vital capacity were also applied (see [16]). Information on age, sex, body mass index (BMI), blood pressure, distance walked, diet and medication use was collected for each participant.

For each individual and each exposure session, three blood samples were collected: two hours before walking, two hours after walking and 24 hours after walking; at the same time TRAP measurements were taken, namely on nitrogen dioxide ($NO_2$), particulate matters $PM_{10}$ and $PM_{2.5}$ and black carbon ($CBLK$). Such measurements are likely to suffer from classical measurement error, as they were collected using a portable size-selective airborne particle sampler and therefore rely on each participant's precision when performing and reporting the experiment, besides being potentially subject to the sources of error in TRAP identified by [12] and reported above. Moreover, as reported in [16], "$NO_2$ concentrations were taken from a stationary monitoring site on Oxford Street repeatedly passed during walks on Oxford Street. Because no monitoring was available in Hyde Park, $NO_2$ concentrations were taken from the nearest representative location sited in a school playground", which suggests the possible presence of a combination of the two error types. Although we do not consider the presence of

Berkson error for the scope of this manuscript, the proposed models can easily be generalized to such a case. Finally, real-time measurements of noise, temperature and relative humidity were obtained at each exposure session. On such samples, different untargeted omic analyses were performed, including metabolomic analyses, on which we focus here. A more detailed description of the study is given in [16], where it is also reported that the study was approved by the UK National Research Ethics Service and that informed written consent was obtained from all participants. [17] conducted analyses to assess the short term metabolomic changes due to exposure to TRAP and identified various associations between air pollution concentrations and levels of metabolites in blood.

## 2.2 The model

The association between metabolite levels and TRAP exposures was assessed in a mixed model framework, using a Bayesian approach and including random effects for the individual, as well as for the location and time point of each measurements. Fixed effects were sex, age, BMI and health group (defined as a categorical variable, with healthy, COPD and IHD as levels), as well as average air pollution concentrations one year before the experiment, used as background or long-term exposure, and instantaneous measurements of the exposure of interest. The four exposures reported above (*CBLK*, $NO_2$, $PM_{25}$ and $PM_{10}$) were considered separately.

The model was formulated as follows:

$$\begin{aligned} m_{ij} &\sim \mathrm{N}(\mu_{ij}, \sigma_e^2) \\ \mu_{ij} &= \alpha + \boldsymbol{\beta}^\top \boldsymbol{X}_{ij} + \beta_{expo} Z_{ij} + \gamma_i + \delta_j, \end{aligned} \tag{1}$$

with $m_{ij}$ being the metabolic feature of individual $i$ at measurement $j$, $Z$ the exposure, $\boldsymbol{X}$ including sex, age, BMI and health group, as well as the annual measurement of the exposure of interest and $\epsilon_{ij} \sim \mathrm{N}(0, \sigma_e^2)$. Random effects were included for the individual $\gamma$ and for the measurement indicator $\delta$, depending on the location and time point of each measurement, for each individual. Normally distributed priors with mean equal to zero and variance equal to 1000 were assigned to the regression coefficients $\boldsymbol{\beta}$, and vague inverse gamma priors with shape and scale equal to 0.01 were assigned to the precisions of the random effects. A graphical representation of the model is reported in Fig 1, while a similar, frequentistic approach to the same model was used in [17]. Given the extreme high dimensionality of the omics dataset we pre-selected only the signals which reported p-values smaller than 0.5 in the frequentist analysis. Our choice of threshold was motivated to reduce the high dimensionality in the data and limit the analyses to a lower number of omic signals, as it would be very impractical to consider all metabolic features at the same time. This value is large enough to avoid false negatives. On the other hand, the hierarchical nature of our analyses leads to shrinkage and reduces the issue of false positives, further justifying the use of a large pvalue. To determine whether there was evidence of an association between the signal and the exposure of interest, we used posterior predictive probabilities to calculate Bayesian $p$-values as tail posterior probabilities. We then corrected for multiple testing using the Bayesian False Discovery Rate (FDR) with level 0.05 [22, 23]. This model, and all models reported in this paper, were formulated using `JAGS` and `INLA` and the code to implement them is reported in the supplementary material. From now on, we will denote this model as the "naive" model, to indicate the fact that it does not account for the presence of measurement error, while we will use "corrected" for the model which includes a component for measurement error.

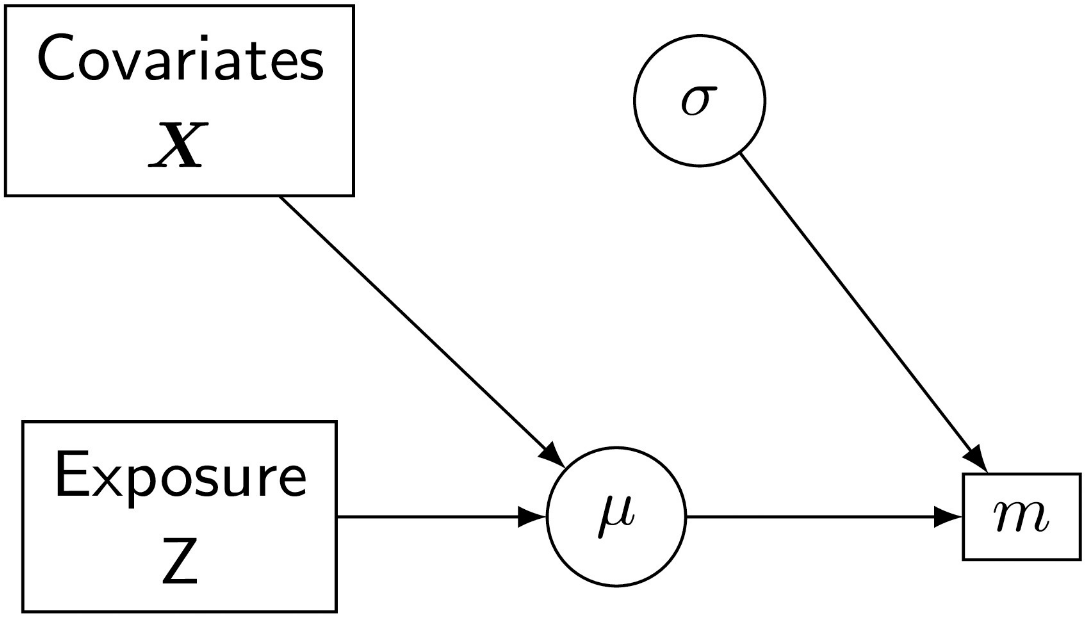

**Fig 1. Bayesian formulation of the basic model to assess association between metabolic levels and TRAP exposure.** Squares are used to indicate data, while circles represent random variables.

https://doi.org/10.1371/journal.pone.0226102.g001

## 3 Classical measurement error

To model the presence of measurement error, we included an additional component in the model by formulating a classical measurement error for the different pollutants, i.e. assuming that the exposure variable $\boldsymbol{Z}$ can be observed only via a proxy $\boldsymbol{W}$, such that

$$\boldsymbol{W} = \boldsymbol{Z} + \boldsymbol{U}, \tag{2}$$

with $\boldsymbol{U} \sim \mathtt{N}(0, \sigma^2_u)$. In the presence of a classical measurement error, an attenuation of the effect of the error-prone variable is expected, as the presence of the additional error variance biases the estimates of the regression parameters towards zero [24]. In fact, bearing in mind that $\hat{\beta}_{expo} = \frac{\sigma(\boldsymbol{W},\boldsymbol{m})}{\sigma^2_W}$, $\sigma^2_W = \sigma^2_Z + \sigma^2_U$, and assuming that the error in $\boldsymbol{W}$ is independent of $\boldsymbol{m}$ and of any other variables, it is quite straightforward to see that the error-prone regression parameter will be given by

$$\hat{\beta}^{\star}_{expo} = \frac{\sigma(\boldsymbol{W},\boldsymbol{m})}{\sigma^2_W} = \frac{\sigma_p(\boldsymbol{Z},\boldsymbol{m})}{\sigma^2_Z + \sigma^2_U} \leq \hat{\beta}_{expo} \ ,$$

and that the quantity that is estimated is $\beta^{\star}_{expo} = \lambda\beta_{expo}$ with $\lambda = \sigma^2_Z/(\sigma^2_Z + \sigma^2_U)$. It is important to underline that, even if an attenuation is usually expected, upward bias is also a possible consequence of classical measurement error even in relatively simple models, due for example to a correlation between covariates [24]. It is therefore necessary to disentangle the error variance from the true variance measured by the proxy in order to obtain unbiased estimates of the regression coefficients. To do so, we formulated a *Bayesian hierarchical model*, as the hierarchical formulation and the possibility to include prior knowledge provide a flexible way to model measurement error [25, 26]. Such formulation was included in the model by adding a latent variable for the exposure, namely a normally distributed variable with mean equal to 0 and variance equal to the error variance. The underlying level of the structure was instead the same as in model (1), resulting in the following hierarchical structure:

$$\begin{aligned} m_{ij} &\sim \mathrm{N}(\mu_{ij}, \sigma_e^2) \\ \mu_{ij} &= \alpha + \boldsymbol{\beta}^{\top}\boldsymbol{X}_{ij} + \beta_{expo}Z_{ij} + \gamma_i + \delta_j \\ W_{ij} &= Z_{ij} + U_{ij} \\ U_{ij} &\sim \mathrm{N}(0, \sigma_U^2). \end{aligned} \tag{3}$$

A vague gamma prior was used for the error variance, with shape and scale parameters equal to 0.01. For all the other variance components and regression coefficients, the same priors were used as in the naive model reported in the previous section.

A graphical representation of the model is reported in Fig 2.

## 4 Multivariate models

### 4.1 Dependency among omic signals

Another source of imprecision and possible bias in the assessment of the association between omic signals and TRAP exposure is potentially given by the formulation of independent models for each omic feature. Dependency across metabolic features is very likely to occur in practice, first of all because 5749 different features are sampled and analysed from the same 60 individuals, and second because they all reflect metabolic pathways and phenomena that are highly correlated in each individual. The flexibility and the layer structure of Bayesian

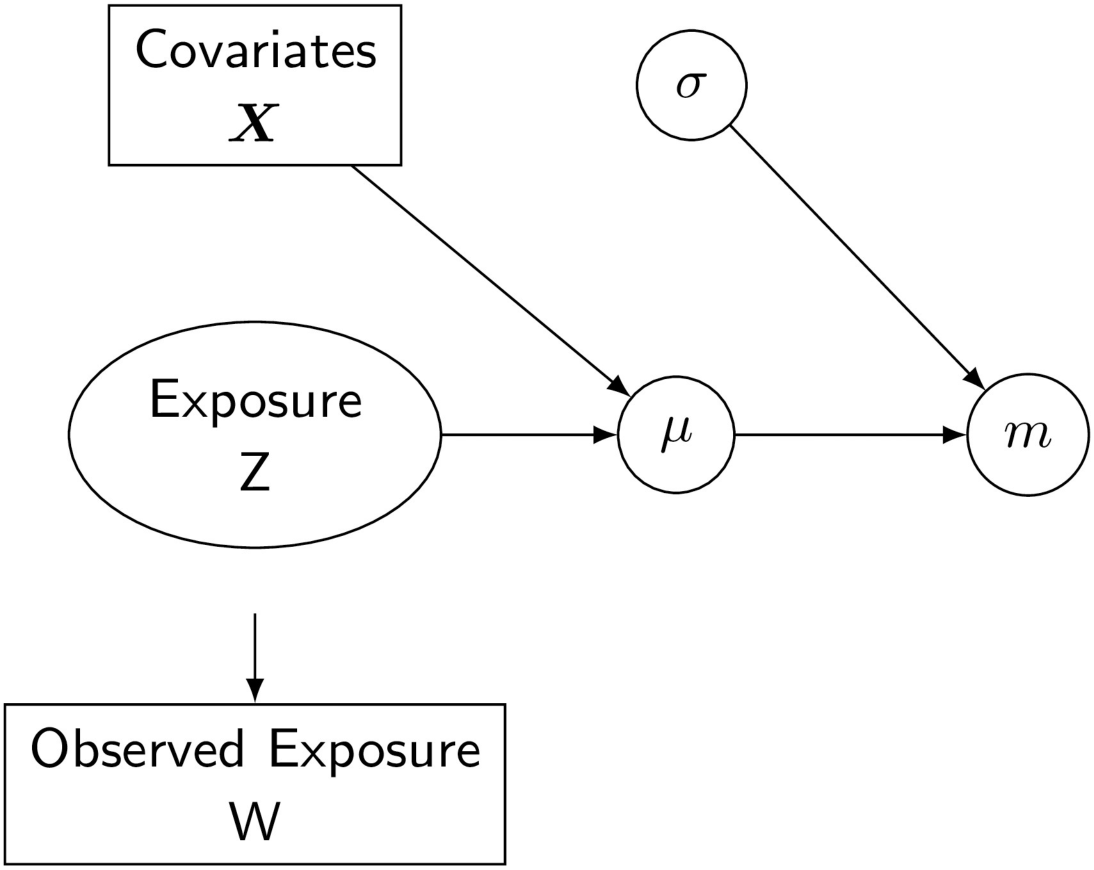

**Fig 2. Bayesian formulation of the classical error model.** *Z* is not observed, *W* is the observed proxy for the exposure.

https://doi.org/10.1371/journal.pone.0226102.g002

hierarchical models makes it straightforward to account for such dependency, namely by using a multivariate response for the omic signals and introducing dependency in their variance covariance structure. The resulting model is a generalization of model (3) and was formulated as follows:

$$\begin{aligned} m_{ij} &\sim \mathrm{N}(\mu_{ij}, \Sigma^2_{e_i}) \\ \mu_{ij} &= \alpha + \beta^\top \mathbf{X}_{ij} + \beta_{expo} Z_{ij} + \gamma_i + \delta_j \\ W_{ij} &= Z_{ij} + U_{ij} \\ U_{ij} &\sim \mathrm{N}(0, \sigma^2_U), \end{aligned} \tag{4}$$

where the response variable followed a multivariate normal distribution $\boldsymbol{m} \sim \boldsymbol{N}(\boldsymbol{\mu}, \Sigma_{\boldsymbol{e}})$, with $\Sigma_{\boldsymbol{e}}$ denoting the covariance matrix of the omics signals and all variances were given a Wishart prior with shape and scale equal to 0.01, in consistency with the univariate model. Note that this formulation with multivariate signals evaluates all the hypoteses at the same time rather than testing each one separately. This implies that model (4) does not require any correction for multiple testing, unlike the initial univariate model, as it naturally accounts for dependency across omics. Therefore, no Bayesian FDR was used in this context and associations were simply assessed by means of posterior predictive distributions.

To model the presence of the error in the exposure variable, we used the same reasoning as before and introduced a latent variable for the exposure, namely a normally distributed variable with mean equal to 0 and variance equal to the error variance. The error structure was again given by (2), following a classical error structure and a vague gamma prior was used for the error variance, with shape and scale parameters equal to 0.01, again in consistency with the formulation and priors used in the univariate model.

A graphical representation of the model is reported in Fig 3.

## 4.2 Dependency among different exposures

Dependency structures are very likely to occur in such studies also between different environmental exposures, as different works in the literature recently pointed out [27, 28]. In particular, all the TRAP measurements show a general common trend based on traffic conditions, as well as weather, humidity and temperature, but also based on other confounding factors registered in each particular day. Moreover, individual-specific factors as activity patterns or time spent in specific locations, as well as respiratory rates can cause a variation in all pollutants at the same time, affecting in our specific case the measurements taken before and after the walk experiment. Such dependency appears to be very frequent between particulate matters of different sizes, as they are induced by similar mechanisms and conditions, besides being measured using the same techniques and instruments, and between PMs and black carbon. The Bayesian hierarchical structures employed in the previous sections can be further generalized to account for such correlation, by modelling more TRAP exposures at the same time. The resulting model was obtained as a generalization of (4) and formulated as follows:

$$\begin{aligned} \mathbf{m}_{ij} &\sim \mathrm{N}(\mu_{ij}, \Sigma^2_{e_i}) \\ \mu_{ij} &= \alpha + \beta^\top \mathbf{X}_{ij} + \beta_{\mathbf{expo}}{}^\top \mathbf{Z}_{ij} + \gamma_i + \delta_j \\ \mathbf{W}_{ij} &= \mathbf{Z}_{ij} + \mathbf{U}_{ij} \\ \mathbf{U}_{ij} &\sim \mathrm{N}(\mathbf{0}, \Sigma^2_U). \end{aligned} \tag{5}$$

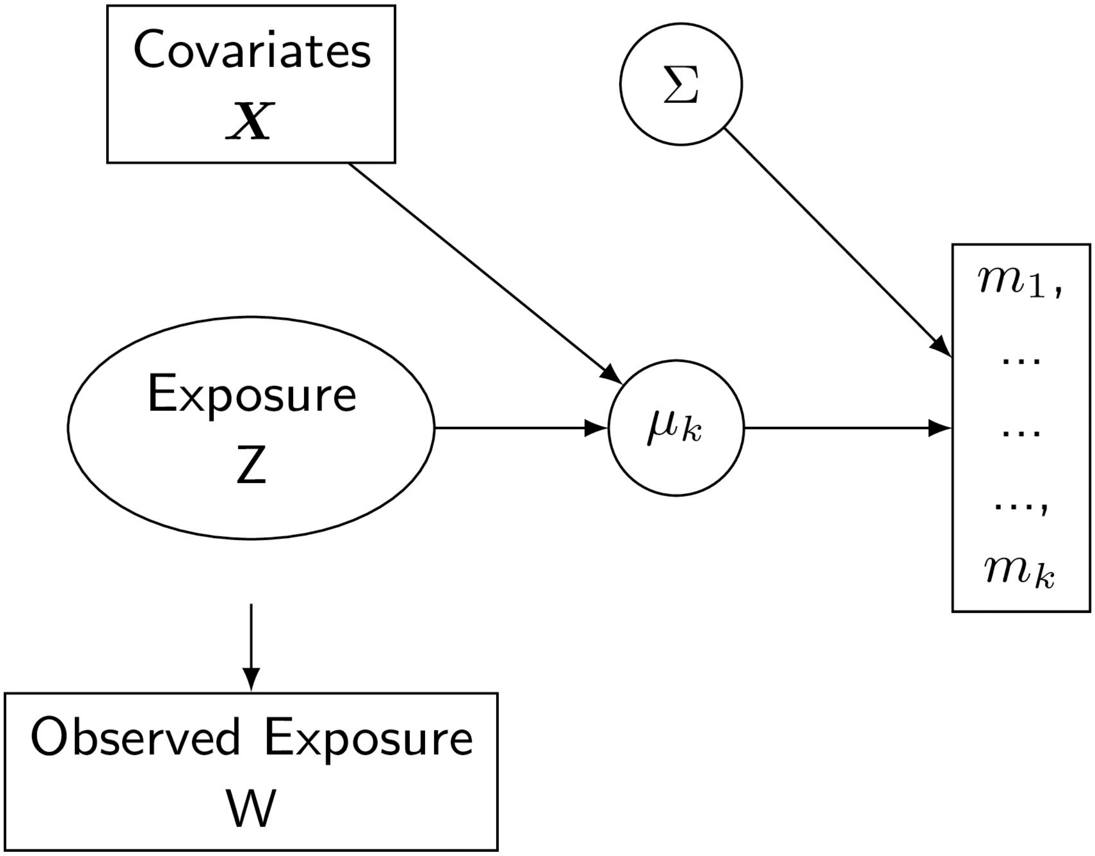

**Fig 3. Bayesian formulation of the multivariate model with classical measurement error.**

https://doi.org/10.1371/journal.pone.0226102.g003

All coefficients and variables were modeled as before and the same priors were given to all of them. Again, the flexible formulation of Bayesian hierarchical structures allows to include as many pollutants as desired and to inform the model on their correlation and variances based on our knowledge. In analogy with the first univariate model, we added a latent variable for the multivariate exposure, following a normal distribution with mean equal to 0 and variance equal to the error variance. To account for dependency among pollutants, *U* was multivariate and modeled as a vector with four dimensions, corresponding to the four different pollutants. The variance covariance matrix of the multivariate exposure, as well as of the multivariate error, was given an uninformative Wishart prior to account for correlation among pollutants.

A graphical representation of the model is reported in Fig 4.

## 5 Implementation

We implemented our models using `JAGS`, as well as `INLA`. In terms of implementation and computation time, `INLA` performed faster than traditional `MCMC` samplers, specifically `JAGS`, but it showed more issues in the matter of space and memory burdens. In particular, when considering dependency among different omic signals, it was possible to implement the error-corrected model in `INLA` with only up to 4 signals, while a substantially higher number of signals were included in the `JAGS` model. In particular, for the single omic model `INLA` took about a quarter of the time employed by `JAGS`. Nevertheless, when including omic-specific measurement error terms, the implementation of more than 4 signals was not possible in

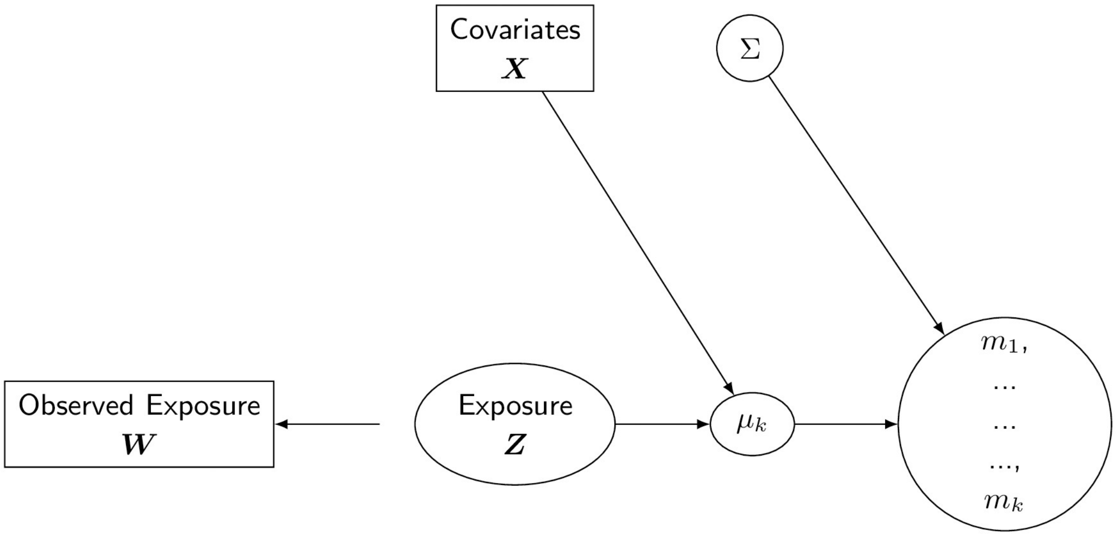
Covariates $\boldsymbol{X}$
$\Sigma$
Observed Exposure $\boldsymbol{W}$
Exposure $\boldsymbol{Z}$
$\mu_k$
$m_1,$
$\dots$
$\dots$
$\dots,$
$m_k$

**Fig 4. Bayesian formulation of the multivariate model with classical measurement error and multivariate exposure.**

https://doi.org/10.1371/journal.pone.0226102.g004

INLA, while it was still feasible in JAGS, although taking very long computational times. These aspects should be taken into account when considering studies with larger sample sizes, where such issues will have even higher impact on the analyses. In this matter, dimension reduction techniques might be employed to select only relevant omic signals, prior to the suggested analyses. All code to reproduce the analyses reported above using both frameworks are provided in the supplementary material.

## 6 Results

The inclusion of a classical measurement error term resulted in different estimates of the association between omic signals and TRAP measurements. Note that the presence of classical measurement error in pollutant measures can cause bias in different directions, and that the effect, as well as the direction of the error correction, is therefore not clear a priori. Fig 5 shows the estimates of regression parameters $\beta_{expo}$ obtained by models (1) and (2) for the selected omics. The error corrected models reported stronger associations between TRAP measurements and metabolomic signals, although with higher uncertainty, which reflects the uncertainty on the error component. While the naive models identified after FDR correction respectively 0, 2, 5 and 24 associations for $CBLK$, $PM_{25}$, $PM_{10}$, $NO_2$, the error corrected models only identified 0, 1, 3 and 5 associations respectively, reflecting the propagation of uncertainty about the error into the posterior distributions of parameters of interest. These signals included a signal from the phenylalanine cluster already found to be associated with $NO_2$ in [17], as well as an unknown signal whose association with $PM_{10}$ was also detected by [17]. This also shows that some associations might be affected by an opposite impact and that attenuation is not the only possible consequence of measurement error. This is also reflected by the width of credible intervals which naturally increases when accounting for measurement error, as a consequence of the uncertainty about this component. When more information is available about the prior distribution of the error, such uncertainty will most likely decrease and potentially smaller credible intervals will be identified.

Such correction acquired a more defined pattern when accounting for the dependency among different omic signals. In such case, the error corrected model retrieved estimates

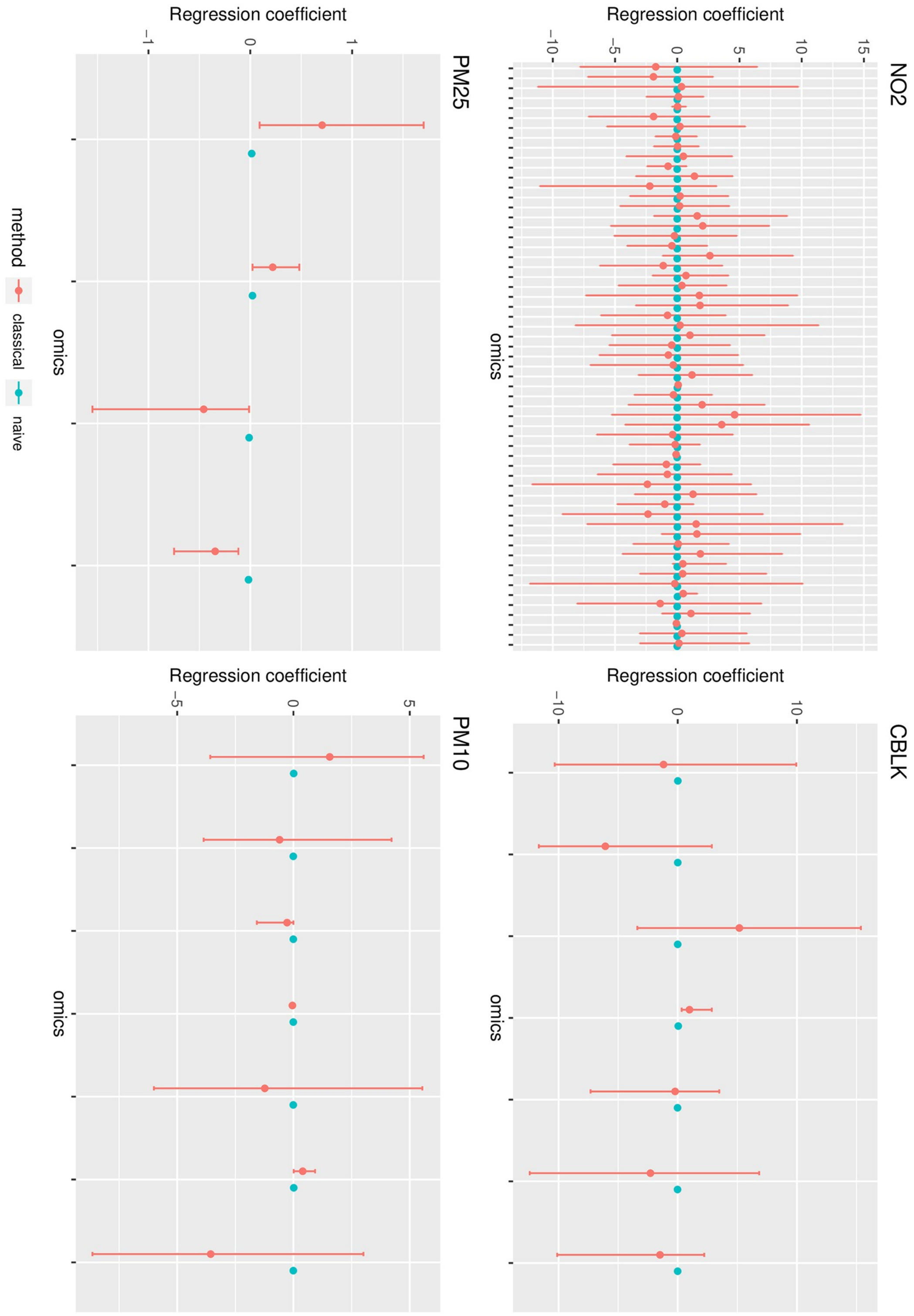

**Fig 5. Regression coefficients with and without classical error modeling in JAGS.** Estimates are reported with their 95% confidence intervals.

https://doi.org/10.1371/journal.pone.0226102.g005

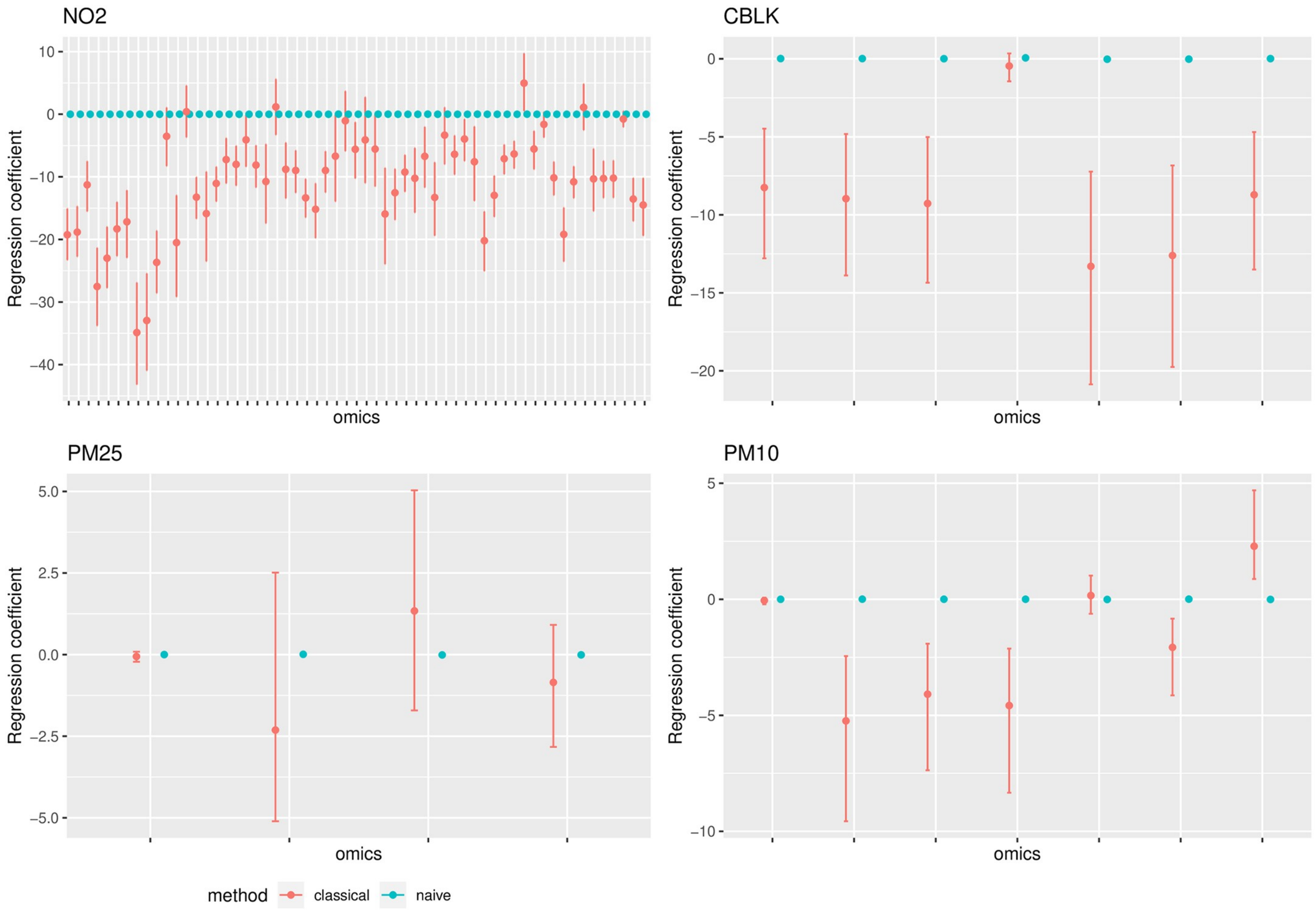

**Fig 6. Regression coefficients with and without classical error modeling in JAGS.** Estimates are reported with their 95% confidence intervals. A correlated structure among omic signals is assumed.

https://doi.org/10.1371/journal.pone.0226102.g006

which were constantly lower than the estimates from the naive models for most omic signals. When comparing the two naive models, with and without accounting for the dependency among signals, the retrieved estimates did not differ substantially. Fig 6 shows the estimates of regression parameters $\beta_{expo}$ obtained by model (4) with and without including a level for the error component. In this case the naive model identified respectively 1, 2, 2 and 4 associations for $CBLK$, $PM_{25}$, $PM_{10}$, $NO_2$, while error corrected model identified respectively 6, 1, 4 and 51 associations. Note that the inclusion of a dependency structure among omic signals and of an error term in the model identifies more associations than the univariate models, because the presence of such dependencies can obscure associations among omic signals and TRAP measurements.

When accounting for dependency among signals and pollutants, all signals that were associated with any of the pollutants were considered. The correction reported by the error corrected models was in most cases higher than in the previous models. Accounting for dependency among pollutants and for correlated measurement errors also retrieved more associations than the other models, leading to higher estimates also between signals and pollutants that were not associated in the first cases. Fig 7 shows the difference in estimates of regression parameters $\beta_{expo}$ obtained by model (5) with and without including a level for the

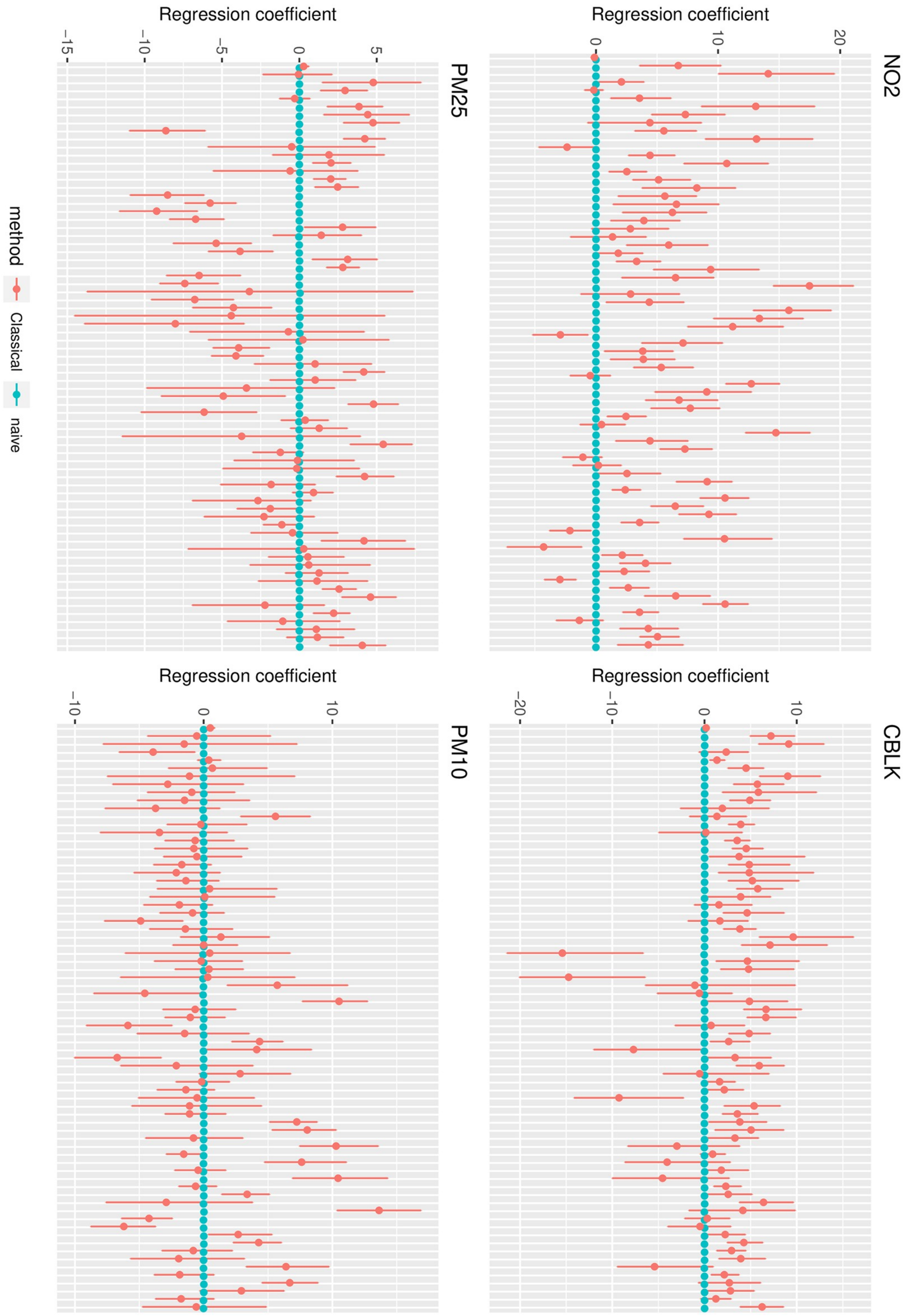

**Fig 7. Regression coefficients with and without classical error modeling in JAGS.** Estimates are reported with their 95% confidence intervals. Dependencies among omic signals and among different TRAP exposures are modeled, as well as dependency among error components on different pollutants in the corrected model. All signals that were associated with any of the pollutants are reported for all pollutants.

https://doi.org/10.1371/journal.pone.0226102.g007

error component. In this case the naive model identified respectively 18, 0, 25 and 16 association for $CBLK$, $PM_{25}$, $PM_{10}$, $NO_2$, while error corrected model identified respectively 7, 1, 22 and 52 associations. This reflects the fact that ignoring the presence of measurement error can obscure associations between signals and pollutants and lead to an underestimation of effect sizes, as well as identifying potentially false associations. On the other hand, models including correlation structures identify even stronger effect sizes when corrected for measurement error, as well as a generally higher number of significant associations between signals and pollutants, which might be obscured when ignoring the presence of bias. In general, more associations were identified when the correlation among omics and among pollutants were modeled, and their number increased when a measurement error term was additionally included in the multivariate models.

## 7 Discussion

We implemented Bayesian hierarchical models to account for the presence of error in measurements of traffic related air pollution. This kind of formulation allows to model several dependency structures in a very flexible way, as well as to include an additional component for measurement error. In our work, we applied such methodology to the study of how TRAP measurements are associated with high-throughput molecular data, namely metabolic features sampled from the exposed individuals in a randomized crossover trial. Our application to the Oxford Street II study showed that the inclusion of a classical error term in the models resulted in corrections of the regression estimates whose extent and direction was not clear a priori, which underlines the importance of explicitly modelling the error component rather than predicting its effect based on prior beliefs. Regression estimates were also corrected by the inclusion of dependency structures in the models, namely dependency among different omic signals and among pollutants. The explicit formulation of such models was possible thanks to the flexible structure of Bayesian hierarchical models, and it was relatively straightforward to embed dependency and measurement error correction in the same hierarchical structure.

This is certainly a major advantage of using Bayesian hierarchical models, which provide a general adaptable way to formulate a broad range of models and structures, as in our case measurement error or dependency structures, and more generally any additional random effect which might be needed in the analysis. Moreover, the use of a Bayesian framework allows to incorporate prior knowledge in the analysis, for example about the error component and parameters, as well as to reflect the prior uncertainty in the posterior distributions of parameters of interest. This requires some knowledge about the error component, in order to properly formulate the measurement error level and to assign reasonable priors. Note that this requirement is not specific to the Bayesian framework, but rather to any error modelling strategies. In fact, it is always necessary to know the error structure (i.e. classical measurement error and its distribution), as well as the error variance, in order to formulate an identifiable error model [29]. In practice, it is not always straightforward to obtain such information, and often assumptions about the error distribution and parameters are vague or potentially incorrect. The advantage of Bayesian models is that the uncertainty in such assumptions can easily be accounted for and propagated to posterior distributions of corrected estimates.

In comparison to general traditional analyses of the association between metabolite levels and air pollution, and specifically to the work of [17], our models allow to account for the presence of measurement error. As a consequence, the estimates of associations between metabolite levels and TRAP exposure are corrected and their absolute values are substantially higher., i.e. they indicate a stronger effect of TRAP on metabolomic features compared with no error adjustment. This practice, and the difference in the results when compared to traditional analyses, show also that ignoring the error can lead to a loss of reproducibility of the studies and to potentially inconsistent results. Moreover, -instead of considering each omic signal separately, we propose to model different signals jointly, so that their correlation is accounted for explicitly. The inclusion of a correlation among exposures is also new with respect to [17]. While traditional analyses are useful to select candidate signals associated with TRAP exposures, the models we suggest in this work explore the texture of this association in more detail, and correct for biases and correlation structures that might distort their quantification and direction. The models we propose use a Bayesian framework to model measurement error and correlation structures based on the data at hand and to incorporate prior beliefs with the evidence given by the observed data. Doing so, it is possible to account for dependencies among different omics and among pollutants, which can sometimes obscure associations between them and which were not taken into consideration in the traditional model, where associations were assessed separately for each omic signal with each pollutant. Moreover, we included a term for measurement error, which was not accounted for in the traditional model and which can also lead to a biased estimation of these associations.

With such framework, it is also possible to treat more general models, for example in terms of error structures. It is indeed sufficient to modify the error level in the hierarchical structure, for instance into a Berkson model. Moreover, given the general formulation of Bayesian hierarchical models, it is possible to include more complex dependency structures, for example spatial correlation between pollutants, which can be easily included in the Bayesian structure, especially with `INLA` [30]. Another possible extension of the method can also be the inclusion of different types of molecular data, which would require a specific and more complex dependency structure among omic signals.

## Supporting information

**S1 Appendix. Code for implementation of all models.** All models are implemented in JAGS in the following files, where "naive" stands for the model with no error correction, "multiomics" for the multivariate model with dependency among omic signals and "multiexpo" for the multivariate model with dependency among exposures.

- multiexpo_classicalME.txt
- multiexpo_naive.txt
- multiomics_classicalME.txt
- multiomics_naive.txt
- univariate_classicalME.txt
- univariate_naive.txt
  (7Z)

## Acknowledgments

This work was supported by the grant FP7 of the European Commission 'Enhanced exposure assessment and omic profiling for high priority environmental exposures in Europe' (EXPOsOMICS grant 308610 to PV). Erica Ponzi was funded by a Doc.Mobility grant from the Swiss National Science Foundation (SNSF grant P1ZHP2_178207).

## Author Contributions

**Conceptualization:** Erica Ponzi, Marta Blangiardo.

**Data curation:** Paolo Vineis, Kian Fan Chung.

**Formal analysis:** Erica Ponzi, Marta Blangiardo.

**Methodology:** Erica Ponzi, Marta Blangiardo.

**Project administration:** Paolo Vineis.

**Software:** Erica Ponzi.

**Supervision:** Paolo Vineis, Marta Blangiardo.

**Validation:** Erica Ponzi, Marta Blangiardo.

**Visualization:** Erica Ponzi.

**Writing – original draft:** Erica Ponzi.

**Writing – review & editing:** Erica Ponzi, Paolo Vineis, Marta Blangiardo.

## References

1. Elliott P, Cuzick J, English D, Stern R. Geographical and Environmental epidemiology: methods for small-area studies. Press OU, editor. New York, NY: Oxford Univ Press; 1992.
2. Pekkanen J, Pearce N. Environmental epidemiology: challenges and opportunities. Environmental Health Perspectives. 2001; 109:1–5. https://doi.org/10.1289/ehp.011091 PMID: 11171517
3. Baker D, Nieuwenhuijsen M. Environmental Epidemiology: Study Methods and Applications. Baker Dean N M, editor. Dean Baker, Mark Nieuwenhuijsen; 2008.
4. Rhomberg LR, Chandalia JK, Long C, Goodman JE. Measurement error in environmental epidemiology and the shape of exposure-response curves. Critical Reviews in Toxicology. 2011; 41(8):651–671. https://doi.org/10.3109/10408444.2011.563420 PMID: 21823979
5. Edwards JK, Keil AP. Measurement error and environmental epidemiology: a policy perspective. Current Environmental Health Reports. 2017; 4(1):79–88. https://doi.org/10.1007/s40572-017-0125-4 PMID: 28138941
6. Shaw PA, Deffner V, Keogh RH, Tooze JA, Dodd KW, Kuechenhoff H, et al. Epidemiologic analyses with error-prone exposures: review of current practice and recommendations. Annals of Epidemiology. 2018; 28:821–828. https://doi.org/10.1016/j.annepidem.2018.09.001 PMID: 30316629
7. Brakenhoff TB, Mitroiu M, Keogh RH, Moons KGM, Groenwold RHH, van Smeden M. Measurement error is often neglected in medical literature: a systematic review. Journal of Clinical Epidemiology. 2018; 98:89–97. https://doi.org/10.1016/j.jclinepi.2018.02.023 PMID: 29522827
8. Mallick B, Hoffman FO, Carroll RJ. Semiparametric regression modeling with mixtures of Berkson and classical error, with application to fallout from the Nevada test site. Biometrics. 2002; 58:13–20. https://doi.org/10.1111/j.0006-341x.2002.00013.x PMID: 11890308
9. Gryparis A, Coull BA, Schwarzt J. Controlling for confounders in the presence of measurement error in hierarchical models; a Bayesian approach. Journal of Exposure Science and Environmental Epidemiology. 2007; 17:S20–S28. https://doi.org/10.1038/sj.jes.7500624 PMID: 18079761
10. Baxter LK, Wright RJ, Paciorek CJ, Laden F, Suh HH, Levy JI. Effects of exposure measurement error in the analysis of health effects from traffic-related air pollution. Journal of Exposure Science and Environmental Epidemiology. 2010; 20:101–11. https://doi.org/10.1038/jes.2009.5 PMID: 19223939
11. Goldman GT, Mulholland JA, Russell AG, Strickland MJ, Klein M, Waller LA, et al. Impact of exposure measurement error in air pollution epidemiology: effect of error type in time-series studies. Environmental Health. 2011; 10(61). https://doi.org/10.1186/1476-069X-10-61 PMID: 21696612
12. 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. Environmental Health Perspectives. 2000; 108:419–426. https://doi.org/10.1289/ehp.00108419 PMID: 10811568
13. 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; 17:377–389. https://doi.org/10.1093/biostatistics/kxv048 PMID: 26621845

**14.** Van Roosbroeck S, Li R, Hoek G, Lebret E, Brunekeef B, Spiegelmann D. Traffic-related outdoor air pollution and respiratory suymptoms in children: the impact of adjustment for exposure measurement error. Epidemiology. 2006; 368:174–184.

**15.** Gryparis A, Zeka A, Schwartz J, Coull BA. Measurement error caused by spatial misalignment in environmental epidemiology. Biostatistics. 2009; 10:258–274. https://doi.org/10.1093/biostatistics/kxn033 PMID: 18927119

**16.** Sinharay R, Gong J, Barratt B, Ohman-Strickland P, Ernst S, Kelly F, et al. Respiratory and cardiovascular responses to walking down a traffic-polluted road compared with walking in a traffic-free area in participants aged 60 years and older with chronic lung or heart disease and age-matched healthy controls: a randomised, crossover study. Lancet. 2017; 391:339–349. https://doi.org/10.1016/S0140-6736(17)32643-0 PMID: 29221643

**17.** van Veldhoven K, Kiss A, Keski-Rahkonen P, Robinot N, Scalbert A, Cullinan P, et al. Impact of short-term traffic-related air pollution on the metabolome—Results from two metabolome-wide experimental studies. Environment International. 2019; 123:124–131. https://doi.org/10.1016/j.envint.2018.11.034. PMID: 30522001

**18.** Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). Journal of the Royal Statistical Society Series B (Statistical Methodology). 2009; 71:319–392.

**19.** Muff S, Riebler A, Held L, Rue H, Saner P. Bayesian Analysis of Measurement Error Models Using Integrated Nested Laplace Approximations. Journal of the Royal Statistical Society- Applied Statistics. 2015; 64:231–252. https://doi.org/10.1111/rssc.12069

**20.** Muff S, Ott M, Braun J, Held L. Bayesian Two-Component Measurement Error Modelling for Survival Analysis Using INLA—A Case Study on Cardiovascular Disease Mortality in Switzerland. Computational Statistics and Data Analysis. 2017; 113:177–193. https://doi.org/10.1016/j.csda.2017.03.001

**21.** Vineis P, Chadeau-Hyam M, Gmuender H, Gulliver J, Herceg Z, Kleinjans J, et al. The exposome in practice: Design of the EXPOsOMICS project. International Journal of Hygiene and Environmental Health. 2017; 220:142–151. https://doi.org/10.1016/j.ijheh.2016.08.001 PMID: 27576363

**22.** Newton MA, Noueiry A, Sarkar D, Ahlquist P. Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics. 2004; 5(2):155–176. https://doi.org/10.1093/biostatistics/5.2.155 PMID: 15054023

**23.** Ventrucci M, Scott EM, Cocchi D. Multiple testing on standard mortality ratios: a Bayesian hierarchical model for FDR estimation. Biostatistics. 2011; 12(1):51–67. https://doi.org/10.1093/biostatistics/kxq040 PMID: 20577014

**24.** Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. Measurement error in nonlinear models, a modern perspective. Boca Raton: Chapman and Hall; 2006.

**25.** Stephens DA, Dellaportas P. Bayesian analysis of generalised linear models with covariate measurement error. In: Bernardo JM, Berger JO, Dawid AP, Smith AFM, editors. Bayesian Statistics 4. Oxford Univ Press; 1992.

**26.** Richardson S, Gilks WR. Conditional independence models for epidemiological studies with covariate measurement error. Statistics in Medicine. 1993; 12:1703–1722. https://doi.org/10.1002/sim.4780121806 PMID: 8248663

**27.** Coker E, Liverani S, Ghosh JK, Jerrett M, Beckermann B, Li A, et al. Multi-pollutant exposure profiles associated with term low birth weight in Los Angeles County. Environment International. 2016; 91:1–13. https://doi.org/10.1016/j.envint.2016.02.011 PMID: 26891269

**28.** Huang G, Lee D amd Scott EM. Multivariate space-time modelling of multiple air pollutants and their health effects accounting for exposure uncertainty. Statistics in Medicine. 2018; 37:1134–1148. https://doi.org/10.1002/sim.7570 PMID: 29205447

**29.** Gustafson P. On Model Expansion, Model Contraction, Identifiability and Prior Information: Two Illustrative Scenarios Involving Mismeasured Variables. Statistical Science. 2005; 20:111–140. https://doi.org/10.1214/088342305000000098

**30.** Blangiardo M, Cameletti M, Baio G, Rue H. Spatial and Spatio-temporal models with R-INLA. Spatial and spatio-temporal epidemiology. 2013; 7:39–55. https://doi.org/10.1016/j.sste.2013.07.003 PMID: 24377114