METHOD ARTICLE

A Method to adjust for measurement error in multiple exposure variables measured with correlated errors in the absence of an internal validation study [version 1; peer review: 1 approved, 1 approved with reservations]

Alexander K. Muoka ^{1,2}, George O. Agogo³, Oscar O. Ngesa², Henry G. Mwambi¹

¹School of Mathematics, Statistics and Computer science, University of Kwazulu-Natal, Pietermaritzburg, South Africa

²Department of Mathematics, Statistics and Physical Sciences, Taita Taveta University, Voi, 635-80300, Kenya

³Centers for Disease Control and Prevention, Nairobi, Kenya

v1 First published: 18 Dec 2020, 9:1486
<https://doi.org/10.12688/f1000research.27892.1>
 Latest published: 18 Dec 2020, 9:1486
<https://doi.org/10.12688/f1000research.27892.1>

Abstract

Difficulty in obtaining the correct measurement for an individual's longterm exposure is a major challenge in epidemiological studies that investigate the association between exposures and health outcomes. Measurement error in an exposure biases the association between the exposure and a disease outcome. Usually, an internal validation study is required to adjust for exposure measurement error; it is challenging if such a study is not available. We propose a general method for adjusting for measurement error where multiple exposures are measured with correlated errors (a multivariate method) and illustrate the method using real data. We compare the results from the multivariate method with those obtained using a method that ignores measurement error (the naive method) and a method that ignores correlations between the errors and true exposures (the univariate method). It is found that ignoring measurement error leads to bias and underestimates the standard error. A sensitivity analysis shows that the magnitude of adjustment in the multivariate method is sensitive to the magnitude of measurement error, sign, and the correlation between the errors. We conclude that the multivariate method can be used to adjust for bias in the outcome-exposure association in a case where multiple exposures are measured with correlated errors in the absence of an internal validation study. The method is also useful in conducting a sensitivity analysis on the magnitude of measurement error and the sign of the error correlation.

Keywords

Measurement error, Internal validation study, Attenuation, Bias, Questionnaire data, Sensitivity analysis, Error correlation

Open Peer Review

Approval Status ? ✓

	1	2
version 1	?	✓
18 Dec 2020	view	view

1. **Erica Ponzi** , University of Oslo, Oslo, Norway

2. **Kesaobaka Molebatsi** , University of Botswana, Notwane Rd, Botswana

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Alexander K. Muoka (alexanderkasyoki@gmail.com)

Author roles: **Muoka AK:** Conceptualization, Data Curation, Formal Analysis, Methodology, Software, Writing – Original Draft Preparation; **Agogo GO:** Conceptualization, Supervision, Writing – Review & Editing; **Ngesa OO:** Conceptualization, Supervision, Writing – Review & Editing; **Mwambi HG:** Conceptualization, Funding Acquisition, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported through a Sub-Saharan Africa Consortium for Advanced Biostatistics training (SSACAB) grant as part of the DELTAS Africa Initiative [107754/Z/15/Z]. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [107754/Z/15/Z] and the UK government. The views expressed in this publication are those of the authors and not necessarily those of AAS, NEPAD Agency, Wellcome Trust or the UK government.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2020 Muoka AK *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Muoka AK, Agogo GO, Ngesa OO and Mwambi HG. **A Method to adjust for measurement error in multiple exposure variables measured with correlated errors in the absence of an internal validation study [version 1; peer review: 1 approved, 1 approved with reservations]** F1000Research 2020, 9:1486 <https://doi.org/10.12688/f1000research.27892.1>

First published: 18 Dec 2020, 9:1486 <https://doi.org/10.12688/f1000research.27892.1>

Abbreviations

HIV: Human immunodeficiency virus; HBCT: Home-based HIV counseling and testing; HSRC: Human sciences research council; NCD: Non-communicable diseases; BMI: Body mass index; kg: kilogram; m^2 : metre squared; g: gram; MCMC: Markov Chain Monte Carlo; CI: Credible interval; JAGS: Just another gibbs sampler; BUGS: Bayesian inference using gibbs sampling; ACF: Autocorrelation function

Introduction

Difficulty in obtaining correct measurements of an individual's long-term exposure is a major challenge in an epidemiological study that investigates the association between a continuous exposure and a health outcome. For instance, several studies estimated the correlations between self-reported intake from a questionnaire and the true long-term intake values to be less than 0.82 for fruits and about 0.72 for vegetables¹⁻⁵, an implication that some of the variation in the diet intake measurements is due to random errors. Due to random error, the association between the dietary intakes and health outcomes may be biased. The effect of measurement error can be quantified using either: (i) the attenuation factor, which quantifies the bias in the association or (ii) the correlation coefficient between the true and the observed exposure, which quantifies the loss of statistical power to detect a significant association (i.e. validity coefficient)⁶.

Validation studies are used to assess the accuracy of the dietary questionnaire⁶⁻¹². A validation study constitutes a small number of individuals from whom dietary intakes are measured repeatedly using an unbiased instrument¹³. There are two types of validation studies: the external and internal validation studies. An internal validation study is conducted on a subset of individuals from the main study, whereas an external validation study is carried on a group of subjects who are not part of the main study, but who are similar in characteristics to individuals in the main study. Validation studies are often expensive to conduct and, in some cases not feasible. Several methods have been proposed to handle measurement error in the absence of internal validation data¹⁴⁻¹⁸.

Agogo *et al.*¹⁴ conducted a sensitivity analysis to investigate the effect of the magnitude of the correlation between errors in the covariates of interest and found that the magnitude of measurement error adjustment is sensitive to the assumed measurement error structure. Dellaportas and Stephens¹⁵ presented a Bayesian method for analysis of non-linear error-in-variable where prior knowledge of the unknown true covariate is incorporated. Huang *et al.*¹⁶ proposed a quantile regression-based non-linear mixed-effects joint models for longitudinal data that simultaneously accounts for a response with non-central location and for covariate with non-normality and measurement error under the Bayesian framework. Lin¹⁷ proposed a Bayesian semi-parametric accelerated failure time model to analyze censored survival data with covariate measurement error and evaluated their method using an intensive simulation study. Muff *et al.*¹⁸ introduced a Bayesian method to handle a mixture of classical and Berkson measurement errors in a single explanatory variable and illustrated their method to studying cardiovascular disease mortality.

The majority of these authors considered a case where one exposure is measured with error (hereafter, a univariate case). In a univariate method, the bias in the association between an outcome and the exposure is adjusted by dividing the unadjusted association estimate by the attenuation factor¹⁹. An attenuation factor is the ratio of the variance of the true exposure to the variance of the observed exposure. This method ignores correlations between the errors, which can lead to substantial bias. In this study, we suggest a general method for adjusting for measurement error where multiple exposures are measured with correlated errors in the absence of an internal validation study (hereafter, a multivariate method). We use real data to illustrate the method in handling a case where three exposures are measured with correlated errors (hereafter, the trivariate method) under a linear regression model and demonstrate the implementation of this method using R software²⁰. Specifically, we use a subset of data from a home-based HIV counseling and testing study that was done in rural and peri-urban communities in KwaZulu-Natal Province, South Africa²¹. We compare the results obtained when using a method that ignores both the measurement error and correlation between the errors (hereafter, a naive method) with those obtained when using univariate and multiple exposures methods. Moreover, we conduct a sensitivity analysis to investigate how the coefficient estimates of parameters of interest are influenced by (1) a change in the level of uncertainty assumed for the limits of the validity coefficients and (2) varying the correlation between errors in the measured exposures.

The remaining sections of this paper are organized as follows. In section 2, we discuss materials and methods used in this study. We present the results of the study in section 3. Finally, we provide a discussion and conclusion in section 4.

Methods

Data and study design

In this work, we use a subset data from a home-based HIV counseling and testing (HBCT) study that was conducted in rural and peri-urban communities in KwaZulu-Natal Province, South Africa, between November 2011 and June 2012²¹. The data were obtained from the Human Sciences Research Council (HSRC) of South Africa²¹. This study was conducted to provide a better understanding of the complexity, severity, and prevalence of non-communicable disease (NCDs) in a community known to have one of the highest rates of HIV incidence and prevalence in the world²¹.

Home-based HIV counseling and testing is a cross-sectional, single-site study in South Africa that aims to increase engagement in HIV care by integrating NCDs screening with community-based HIV testing²². A random sampling approach was used, where 587 participants over the age of 18 were selected from 50,000 people living in the Mpumzu suburb²¹. Anthropometric and biological measures were collected in the survey with the purpose of establishing the prevalence of a range of NCDs and associated risk factors. Eligible individuals participated in a face-to-face interview, physical, psychological and clinical examinations. Persons younger than 18 years living in Mpumzu and all household members not previously enrolled, and members unable to give written consent were excluded from the study. Mobile phones were used for data collection to increase efficiency in data capture and analysis²¹.

In our study, we used a subset data consisting of 76 individuals who self-reported the number of cigarettes smoked, fruit and vegetable consumption. We use the dataset to illustrate the multivariate method in modeling the amount of association between body mass index (BMI) and three exposures (smoking, fruit, and vegetable intakes). BMI was measured in kg/m^2 , while smoking was measured as the average number of cigarettes smoked per day. Initially, fruit and vegetable intakes were measured in terms of the number of servings consumed per day. It is often assumed that a standard portion of fruit/vegetable weighs about 80g⁵. Therefore, for this study, we converted the number of servings to grams per day (g/day) by multiplying the reported number of servings by 80g. The subset data has the following three properties that make it suitable for use in this work: (1) measurement error in the recorded number of cigarettes smoked due to possible misreporting, (2) measurement error in fruits and vegetable consumption due to recall bias, and conversion of the number of servings of fruits and vegetables into grams, and (3) the measurement error in the three exposures is often correlated, for instance, smokers are likely to over-report fruit and vegetable intakes due to their beneficial effects, and to under-report the number of cigarettes they smoke due to the associated harmful effects. Epidemiologically, BMI is used as a risk factor of a health outcome. However, in this study, we model BMI as an outcome as in other several studies, for instance,^{23–26}. The subset data is only used to illustrate the method and not to draw inference.

Ethical statement

Ethics approval was granted by both HSRC Research Ethics Committee (REC: 1/26/05/11) and the University of Washington Institutional Review Board (48733). Informed written consent was obtained from each participant in the study. Participants were provided with written information on the study (including the study's background and objectives) and their rights regarding participation and withdrawal at any time.

A measurement error model for the data

An interest in epidemiological study could be to investigate the association between BMI and three exposures namely: fruit, vegetable and smoking using the multiple linear regression

$$Y = \beta_0 + \beta_{X_1}X_1 + \beta_{X_2}X_2 + \beta_{X_3}X_3 + \epsilon, \quad (1)$$

where Y denotes the BMI, β_0 is the intercept, β_{X_1} , β_{X_2} and β_{X_3} are the coefficient parameters for the true long-term fruit (X_1), vegetable (X_2) and cigarette (X_3) intake respectively and ϵ is the random error term. In this study, we use vegetable intake and cigarette smoking as confounders and assume that the main interest is in estimating β_{X_1} . In practice, the true intakes are unobservable and, therefore, the intakes recorded in self-reported questionnaires are used. Let W_1 , W_2 and W_3 denote the measured versions of X_1 , X_2 and X_3 , respectively. The use of W_p 's in place of X_p 's, ($p = 1, 2, 3$), in Equation (1) yields biased estimates $\hat{\beta}_{W_1}$, $\hat{\beta}_{W_2}$ and $\hat{\beta}_{W_3}$ of β_{X_1} , β_{X_2} and β_{X_3} respectively. Let $\hat{\beta}_W = (\hat{\beta}_{W_1}, \hat{\beta}_{W_2}, \hat{\beta}_{W_3})^T$.

We assumed that the observed exposures are related to the true exposures with additive measurement error as

$$W_i = \alpha_{0i} + \alpha_{1i}X_i + \epsilon_{W_i}, \quad i = 1, 2, 3 \quad (2)$$

where $\epsilon_w = (\epsilon_{w_1}, \epsilon_{w_2}, \epsilon_{w_3})^T$, $\epsilon_w \sim N(0, \Sigma_{\epsilon_w})$; $\mathbf{W} = (W_1, W_2, W_3)^T$; $\alpha_0 = (\alpha_{01}, \alpha_{02}, \alpha_{03})^T$, $\alpha_1 = (\alpha_{11}, \alpha_{12}, \alpha_{13})^T$; with the terms in α_0 and α_1 quantifying the constant bias and the proportional scaling bias respectively; ϵ_w is a random error term, ϵ_{w_i} is assumed to be independent of the true exposure X_i and the systematic bias components, α_{0i} and α_{1i} .

Bias adjustment methods

A univariate method. In a univariate case, the bias in the association between an outcome and an exposure is adjusted by dividing the unadjusted association estimate by the attenuation factor¹⁹. Attenuation factor (λ_i) is defined as $\lambda_i = \text{var}(X_i)/\text{var}(W_i)$, i.e., the ratio of the variance of the true exposure to the variance of the observed exposure, also referred to as reliability ratio. This method ignores correlations between the errors and also the correlation between the true exposures.

Multivariate method. We propose and describe a general approach for handling p-exposures ($p \geq 2$) measured with correlated errors. For simplicity and without loss of generality, we assume that W_i is measured without systematic bias (i.e., $\alpha_{0i} = 0$, $\alpha_{1i} = 1$ in Equation 2). For multiple exposures measured with correlated errors, the adjusted association estimates can be obtained by pre-multiplying the unadjusted association estimates by the inverse of the transpose of attenuation-contamination matrix as

$$\hat{\beta}_X^* = (\hat{\Lambda}_p^T)^{-1} \hat{\beta}_W^*, \quad (3)$$

where $\hat{\beta}_X^*$ and $\hat{\beta}_W^*$ denotes vectors of true and biased coefficients for the p-exposures respectively and Λ_p denotes a $p \times p$ attenuation-contamination matrix^{19,27}. The off-diagonal elements in Λ are known as contamination factors while the diagonal elements are called attenuation factors¹⁴. Noteworthy, the attenuation factor quantifies the bias in the association between an outcome and an exposure. In contrast, the contamination factor quantifies the effect of measurement error in one exposure variable on the other exposure variable's estimate. $\hat{\beta}_W^*$ in Equation (3) can be obtained from the observed questionnaire data.

In the multiple exposures case, the estimate of attenuation-contamination matrix $\hat{\Lambda}_p$ is defined as

$$\hat{\Lambda}_p = \underbrace{\begin{bmatrix} \hat{\sigma}_{X_1}^2 & \hat{\sigma}_{X_1 X_2} & \cdots & \hat{\sigma}_{X_1 X_p} \\ \hat{\sigma}_{X_2 X_1} & \hat{\sigma}_{X_2}^2 & \cdots & \hat{\sigma}_{X_2 X_p} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\sigma}_{X_p X_1} & \hat{\sigma}_{X_p X_2} & \cdots & \hat{\sigma}_{X_p}^2 \end{bmatrix}}_{\hat{\Sigma}_X^*} \underbrace{\begin{bmatrix} \hat{\sigma}_{W_1}^2 & \hat{\sigma}_{W_1 W_2} & \cdots & \hat{\sigma}_{W_1 W_p} \\ \hat{\sigma}_{W_2 W_1} & \hat{\sigma}_{W_2}^2 & \cdots & \hat{\sigma}_{W_2 W_p} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\sigma}_{W_p W_1} & \hat{\sigma}_{W_p W_2} & \cdots & \hat{\sigma}_{W_p}^2 \end{bmatrix}}_{\hat{\Sigma}_W^{*-1}}^{-1}, \quad (4)$$

where $\hat{\Sigma}_X^*$ is the estimate of covariance matrix of the true exposures, $\hat{\Sigma}_W^{*-1}$ is the inverse of the estimate of covariance matrix of the measured exposures, $\hat{\sigma}_{X_i}^2$ is the variance estimate of X_i ($i = 1, 2, \dots, p$); $\hat{\sigma}_{X_i X_j}$ ($j = 1, 2, \dots, p$; $i \neq j$) denotes the covariance estimate between the true exposures; $\hat{\sigma}_{W_i}^2$ is the variance estimate of W_i ; $\hat{\sigma}_{W_i W_j}$ ($i \neq j$) is the covariance estimate between the observed exposures.

The elements of the variance-covariance matrix of the observed exposures, Σ_W^* , are estimated from the observed data. The variances of the true exposures, $\sigma_{X_i}^2$'s, can be estimated using validity coefficients for the questionnaire. According to Kipnis *et al.*⁶, the validity coefficient is given by:

$$\begin{aligned} \rho_{W_i X_i} &= \frac{\text{cov}(W_i, X_i)}{\sqrt{\text{var}(W_i)\text{var}(X_i)}}, \quad i = 1, 2, \dots, p. \\ &= \frac{\sigma_{X_i}}{\sigma_{W_i}}, \end{aligned} \quad (5)$$

where W_i is assumed to be the measured with error term only and ϵ_{w_i} is assumed to be independent of X_i . From Equation (5), we estimate the variance of the true exposures as

$$\hat{\sigma}_{X_i}^2 = (\hat{\rho}_{W_i X_i} \hat{\sigma}_{W_i})^2, \quad (6)$$

by incorporating external validation information on $\rho_{w_i X_i}$. To obtain covariances between the true exposures, one of the following two approaches is used: (i) if external information about the correlation between true exposures (i.e. $\hat{\rho}_{X_i X_j}$) is available, we obtain covariances between true exposures as follows:

$$\hat{\sigma}_{X_i X_j} = \hat{\rho}_{X_i X_j} \hat{\sigma}_{X_i} \hat{\sigma}_{X_j}, \quad i \neq j, \quad (7)$$

where $\hat{\sigma}_{X_i}$ are obtained as shown in Equation (6); (ii) if we can obtain prior information about the correlation between the errors in the observed exposures, $\hat{\rho}_{\epsilon_{w_i} \epsilon_{w_j}}$, we can solve for $\hat{\sigma}_{X_i X_j}$ by decomposing the covariance of observed exposures into unknown covariance between true exposures and unknown covariance between errors as follows:

$$\begin{aligned} \hat{\sigma}_{w_i w_j} &= \hat{\sigma}_{X_i X_j} + \hat{\sigma}_{\epsilon_{w_i} \epsilon_{w_j}} + \underbrace{\hat{\sigma}_{X_i \epsilon_{w_j}}}_{0} + \underbrace{\sigma_{X_j \epsilon_{w_i}}}_{0} \\ &= \hat{\sigma}_{X_i X_j} + \hat{\rho}_{\epsilon_{w_i} \epsilon_{w_j}} \hat{\sigma}_{\epsilon_{w_i}} \hat{\sigma}_{\epsilon_{w_j}}, \end{aligned} \quad (8)$$

where X_i and ϵ_{w_j} , X_j and ϵ_{w_i} are assumed to be uncorrelated.

From Equation (2) and Equation (6), the estimate of the error variance $\hat{\sigma}_{\epsilon_{w_i}}^2$ is

$$\begin{aligned} \hat{\sigma}_{\epsilon_{w_i}}^2 &= \hat{\sigma}_{w_i}^2 - \underbrace{\hat{\sigma}_{w_i}^2 \hat{\rho}_{w_i X_i}^2}_{\hat{\sigma}_{X_i}^2} \\ &= \hat{\sigma}_{w_i}^2 (1 - \hat{\rho}_{w_i X_i}^2), \end{aligned} \quad (9)$$

See Appendix B of the extended data²⁸ for the proof.

From Equation (8)–Equation (9), the covariances between the true exposures are given by

$$\hat{\sigma}_{X_i X_j} = \hat{\sigma}_{w_i w_j} - \hat{\rho}_{\epsilon_{w_i} \epsilon_{w_j}} \hat{\sigma}_{w_i} \hat{\sigma}_{w_j} \sqrt{(1 - \hat{\rho}_{w_i X_i}^2)(1 - \hat{\rho}_{w_j X_j}^2)}, \quad (i \neq j), \quad (10)$$

Using the observed data and external information, we can determine all the terms required to estimate the attenuation-contamination matrix, Λ , as shown in Equation (4) and adjust for the bias in the association between the exposures measured with error and the outcome using Equation (3).

Illustration of the multivariate method using the study data

We illustrate a method that accounts for uncertainty in the validity measures attributable to heterogeneity in the study populations and in parameter estimation. The proposed Bayesian method applies Markov Chain Monte Carlo (MCMC) estimation approach to combine observed self-reported data and external validation data in adjusting for measurement error in three exposures measured with correlated errors. MCMC is a class of algorithms that samples from the posterior distributions by traversing the parameter space²⁹. The posterior distribution is obtained by updating the prior distribution with observed data. The steps for implementing the trivariate method are described below.

We first obtained external information on validity coefficients and generated validity coefficients for use by interpreting the lower and upper limits obtained from the literature as the 95% credible intervals (CIs) of the distribution of possible values respectively. Due to the skewed distribution of validity coefficients, Fisher's transformation was used to generate the validity coefficients as explained in the next section.

Second, for the observed exposures, we estimated the posterior distribution of the covariance matrix (Σ_w). The exposures were assumed to follow a multivariate normal distribution with mean and covariance, i.e., $\mathbf{W} \sim N_3(\mu_w, \Sigma_w)$. We assumed a weakly informative multivariate normal prior for μ_w as $\mu_{w \text{ prior}} \sim N_3(0, 10^6 \mathbf{I}_3)$, where $\mathbf{I}_3$ is a 3×3 identity matrix. In a multivariate normal distribution, Σ_w must satisfy two conditions: (1) be positive definite (i.e. $\mathbf{W}^T \Sigma_w \mathbf{W} > 0$, for all $\mathbf{W}$) and (2) be a symmetric matrix. The semi-conjugate prior distribution for Σ_w , which has these two properties, is the inverse-Wishart distribution²⁹. To minimize the influence of the prior information on the estimate of Σ_w , we considered weakly informative inverse-wishart prior as $\Sigma_{w \text{ prior}} \sim IW(\mathbf{I}_3, \nu)$, where $\nu = 3$ is the degrees of freedom.

Third, using the validity coefficients generated from the external data and the posterior distribution of covariance matrix for observed exposures, we estimated the variance of true intakes, $\hat{\sigma}_{X_i}^2$ ($i = 1, 2, 3$), using the relationship given in Equation (6) so that

$$\begin{aligned}\hat{\sigma}_{X_1}^2 &= (\hat{\rho}_{W_1 X_1} \hat{\sigma}_{W_1})^2 \\ \hat{\sigma}_{X_2}^2 &= (\hat{\rho}_{W_2 X_2} \hat{\sigma}_{W_2})^2 \\ \hat{\sigma}_{X_3}^2 &= (\hat{\rho}_{W_3 X_3} \hat{\sigma}_{W_3})^2.\end{aligned}\quad (11)$$

The covariances between true intakes ($\hat{\sigma}_{X_i X_j}$; $j = 1, 2, 3$) were estimated as,

$$\begin{aligned}\hat{\sigma}_{X_1 X_2} &= \hat{\sigma}_{W_1 W_2} - \hat{\rho}_{\epsilon_{W_1} \epsilon_{W_2}} \hat{\sigma}_{W_1} \hat{\sigma}_{W_2} \sqrt{(1 - \hat{\rho}_{W_1 X_1}^2)(1 - \hat{\rho}_{W_2 X_2}^2)} \\ \hat{\sigma}_{X_1 X_3} &= \hat{\sigma}_{W_1 W_3} - \hat{\rho}_{\epsilon_{W_1} \epsilon_{W_3}} \hat{\sigma}_{W_1} \hat{\sigma}_{W_3} \sqrt{(1 - \hat{\rho}_{W_1 X_1}^2)(1 - \hat{\rho}_{W_3 X_3}^2)} \\ \hat{\sigma}_{X_2 X_3} &= \hat{\sigma}_{W_2 W_3} - \hat{\rho}_{\epsilon_{W_2} \epsilon_{W_3}} \hat{\sigma}_{W_2} \hat{\sigma}_{W_3} \sqrt{(1 - \hat{\rho}_{W_2 X_2}^2)(1 - \hat{\rho}_{W_3 X_3}^2)}.\end{aligned}\quad (12)$$

by incorporating external validation information on correlation between the errors ($\rho_{\epsilon_{W_i} \epsilon_{W_j}}$). We generated the correlation between errors from a plausible range guided by correlation in the observed data and prior expert information on the most likely sign of the correlation between the exposures, as described in the next section.

Having obtained the covariance matrices of the true and observed exposures, we estimated the attenuation-contamination matrix (Λ_3) from their joint distribution as

$$\hat{\Lambda}_3 = \underbrace{\begin{bmatrix} \hat{\sigma}_{X_1}^2 & \hat{\sigma}_{X_1 X_2} & \hat{\sigma}_{X_1 X_3} \\ \hat{\sigma}_{X_1 X_2} & \hat{\sigma}_{X_2}^2 & \hat{\sigma}_{X_2 X_3} \\ \hat{\sigma}_{X_1 X_3} & \hat{\sigma}_{X_2 X_3} & \hat{\sigma}_{X_3}^2 \end{bmatrix}}_{\hat{\Sigma}_X} \underbrace{\begin{bmatrix} \hat{\sigma}_{W_1}^2 & \hat{\sigma}_{W_1 W_2} & \hat{\sigma}_{W_1 W_3} \\ \hat{\sigma}_{W_1 W_2} & \hat{\sigma}_{W_2}^2 & \hat{\sigma}_{W_2 W_3} \\ \hat{\sigma}_{W_1 W_3} & \hat{\sigma}_{W_2 W_3} & \hat{\sigma}_{W_3}^2 \end{bmatrix}}_{\hat{\Sigma}_W^{-1}}^{-1}, \quad (13)$$

where $\hat{\Sigma}_X$ is the estimate of covariance matrix of the three true exposures, $\hat{\Sigma}_W^{-1}$ is the inverse of the estimate of covariance matrix of the three measured-with-error exposures, $\hat{\sigma}_{X_i}^2$ is the variance estimate of X_i ($i = 1, 2, 3$); $\hat{\sigma}_{X_i X_j}$ ($i \neq j$) denotes the covariance estimate between the true exposures; $\hat{\sigma}_{W_i}^2$ is the variance estimate of W_i ; $\hat{\sigma}_{W_i W_j}$ ($i \neq j$) is the covariance estimate between the observed exposures.

Lastly, we fitted a Bayesian multiple linear regression model (hereafter, naive method) to obtain the posterior distributions of the unadjusted coefficient estimates $\hat{\beta}_W = (\hat{\beta}_{W_1}, \hat{\beta}_{W_2}, \hat{\beta}_{W_3})^T$. In the naive model, we assumed weakly informative normal independent priors by choosing a very small precision (large variance) for the unadjusted coefficient estimates as $\beta_{W_i \text{ prior}} \sim N(0, 10^6)$. The adjusted coefficient estimates $\hat{\beta}_X$ were then obtained from the joint posterior distribution of $\hat{\Lambda}_3$ and $\hat{\beta}_W$ as

$$\hat{\beta}_X = (\hat{\Lambda}_3^T)^{-1} \hat{\beta}_W. \quad (14)$$

Software implementation of the trivariate method

We implemented the trivariate method in R version 3.6.3 using **rjags** (version 4-10), **coda** (version 0.19-3), **MCMCpack** (version 1.4-9), and **mvtnorm** (version 1.1-1) packages. To facilitate Bayesian estimation of the covariance matrix of the observed exposures (Σ_W), **rjags** package was used to provide an interface from R to the JAGS library³⁰. JAGS is a gibbs sampler that uses MCMC to draw dependent samples from the posterior distribution of the parameters³¹. The Bayesian estimation of Σ_W proceeded in the following steps: (1) defining a model for Σ_W under Bayesian inference using gibbs sampling (BUGS) algorithm in a stand alone file, (2) reading the model file using the `jags.model` function, (3) updating the model using the `update` method for jags objects and (4) extracting the posterior samples of the model using the `coda.samples` function from the coda package.

MCMCregress function from the **MCMCpack** package was used to generate a posterior density sample from the naive linear regression model³². MCMC convergence diagnostics of all the model parameters was done using

trace plots and autocorrelation (ACF) plots from the `coda` package³³. See extended data: Appendix C²⁸ for convergence diagnostics results. For each model, the burn-in iterations were set to 2,000 and 10,000 MCMC iterations were run after the burn-in iterations. Every first sample value was kept in the MCMC simulations by using a thinning interval of 1. When compiling a JAGS model, an initial sampling step may be needed during which the samplers learn their behaviour to maximize their performance³⁴. Therefore, the number of iterations for adaptation in the the jags model was set to 500. The results were presented in terms of density plots, posterior mean and median. We compared the results obtained under naive, univariate, and trivariate methods. The R code used for analysis is presented in the extended data²⁸.

External information on the validity coefficient and error correlations for the study data

External information on the validity coefficient and error correlations for fruit, vegetable, and cigarette information was obtained from the literature. According to Kaaks *et al.*¹, the validity coefficient of self-reported fruit intake ranges from 0.33 to 0.79, while that of vegetable intake ranged from 0.30 to 0.60. A meta-analysis study on the validity of questionnaires assessing fruit and vegetable consumption by Collese *et al.*² reported validity coefficients of 0.26 for vegetables and 0.49 for fruits. Other similar validation studies reported validity coefficients in the aforementioned ranges for fruits and vegetables^{3,4,35}. Therefore, based on these information we considered a range of 0.3 to 0.8 for fruits and a range of 0.25 and 0.7 for vegetables.

In the Scottish Heart Health Study of 2,849 men and 2,900 women³⁶, the correlation between the self-reported number of cigarettes and biochemical measures was reported between 0.67 and 0.72. In a study on the validation of self-reported smoking by analysis of hair for nicotine and cotinine³⁷, the validity coefficient between the number of cigarettes smoked per day and nicotine/cotinine levels in hair and plasma was found to be between 0.48 and 0.63, while the correlation between the average number of cigarettes smoked and carboxyhemoglobin was 0.70. In a follow-up study to examine the relationships among self-reported cigarette consumption, exhaled carbon monoxide, and urinary cotinine/creatinine ratio in pregnant women³⁸, a validity coefficient in the range of 0.61 to 0.70 was reported. A study by Stram *et al.*³⁹ found the correlation between the self-reported number of cigarettes smoked and the true lung dose to be between 0.40 and 0.70, and this range was consistent with the findings from the previously discussed related validation studies. Based on this information, we considered a validity coefficient range of 0.40 and 0.70.

We generated the correlation between errors from plausible ranges that were determined based on the correlation in the observed data and the most probable sign of the correlation among fruits, vegetables, and cigarettes as explained below:

- Since the correlation coefficient between fruit and vegetable intake in the observed data was positive, we also assumed the error correlation between fruit and vegetables to be mostly positive;
- An investigation on the correlation coefficient between cigarette smoking and fruits/vegetable intake in the observed data showed a negative correlation coefficient. Based on this and the fact that persons who tend to overstate fruit and vegetable consumption are likely to understate the number of cigarettes smoked, we assumed the error correlation to be mostly negative.

We obtained the upper limits of error correlations by assuming that the error covariance equals the covariance in the observed data and set the lower limit of the error correlation to zero, based on the assumption that the covariance in the observed data equals the covariance between the true intakes¹⁴.

Estimating the distribution of $\rho_{w_i x_i}$

Using the range of plausible values obtained from external validation information, we generated the validity coefficients using the Fisher-Z transformation method by assuming that the reported lower and upper limits are 0.05 and 0.95 quantiles of the uncertainty distribution, respectively. Fisher Z-transformation is a commonly used method to transform the sampling distribution of correlation coefficients to become approximately normally distributed^{40,41}. The procedure is as outlined below:

- Using the Fisher Z-transformation formula

$$F_{Z_i} = 0.5 \left[\ln(1 + \rho_{w_i x_i}) - \ln(1 - \rho_{w_i x_i}) \right], \quad (15)$$

transform the lower (r_l) and upper (r_u) limits of the validity coefficient $\rho_{w_i x_i}$ to get the corresponding Fisher-Z transformed values F_{Z_l} and F_{Z_u} respectively.

- (ii) Compute the mean μ_{Z_i} and the standard deviation σ_{Z_i} of F_{Z_i} as $\mu_{Z_i} = 0.5(F_{Z_u} - F_{Z_l})$ and $\sigma_{Z_i} = \frac{0.5(F_{Z_u} - F_{Z_l})}{Z_{\alpha/2}}$ where $Z_{\alpha/2}$ is the $(1 - \frac{\alpha}{2})\%$ quantile of a standard normal random variable.
- (iii) Generate F_{Z_i} 's as $F_{Z_i} \sim N(\mu_{Z_i}, \sigma_{Z_i}^2)$
- (iv) Using the inverse of Fisher Z-transformation, back-transform the generated F_{Z_i} 's to validity coefficient as

$$\rho_{W_i X_i} = \frac{\exp(2F_{Z_i}) - 1}{\exp(2F_{Z_i}) + 1}. \quad (16)$$

Sensitivity analysis

We investigated how varying the level of uncertainty assumed for the limits of the validity coefficients reported from literature affected the estimates for fruit, vegetable, and the average number of cigarettes smoked. We also investigated how the estimates varied with the magnitude of the correlation between errors in fruit and vegetable intake, fruit and cigarette smoking, and vegetable and cigarette smoking. This helps determine the estimates' sensitivity to various magnitudes of CI and the correlation between errors when using the multivariate method.

Results

Table 1 presents regression coefficients estimates for fruit intake (g/day), vegetable intake (g/day), and the average amount of smoked cigarettes a day obtained using the naive method and the two bias adjustment methods (i.e., univariate and trivariate methods). The regression coefficient estimate adjusted for bias using either the univariate or trivariate method was greater in absolute value than that obtained using the naive method. Specifically, for fruit intake and the average number of cigarettes smoked, the bias-adjusted coefficient estimates were three times as large as the naive coefficient estimates. For vegetable intake, the increase in the strength of the association was about four times as compared to the naive regression coefficient estimates.

For both fruit intake and the average number of cigarettes smoked, the univariate method gave slightly greater estimates while the bias-adjusted values for vegetable intake were slightly lower in the univariate method. The variability of the regression coefficient estimate of the number of cigarettes smoked was higher than that for both fruits and vegetable intake. Again, the variability in either the univariate or trivariate method was higher than in the naive method due to uncertainty involved in adjusting for measurement error.

Figure 1–Figure 3 show the kernel densities representing the distributions of adjusted for measurement error (solid curves) and naive (dotted curves) estimates for fruits intake, vegetable intake, and the number of cigarettes smoked,

Table 1. Comparison of posterior Mean (Standard Deviation) and posterior Median for the estimates of fruit (g/day), vegetable(g/day) and average number of cigarettes smoked per day unadjusted for measurement error (naive estimates) and adjusted for measurement error using univariate and trivariate methods.

Method	Estimate for fruit intake		Estimate for vegetable intake	
	Mean (SD)	Median	Mean (SD)	Median
Naive	0.009 (0.012)	0.009	0.008 (0.014)	0.008
Univariate	0.026 (0.036)	0.027	0.031 (0.051)	0.031
Trivariate	0.026 (0.036)	0.026	0.033 (0.051)	0.033
Method	Estimate for smoking			
	Mean (SD)	Median		
Naive	-0.253 (0.640)	-0.247		
Univariate	-0.740 (1.874)	-0.721		
Trivariate	-0.714 (1.875)	-0.695

Figure 1. Kernel densities for the distribution of adjusted for measurement error and unadjusted estimates for fruit intake. The solid vertical lines show the posterior means of coefficient estimates adjusted for bias; the dotted vertical lines indicate the posterior means of unadjusted coefficient estimates.

Figure 2. Kernel densities for the distribution of adjusted for measurement error and unadjusted estimates for vegetable intake.

respectively. The solid vertical lines on the density plots depict the posterior mean of the adjusted regression coefficients, while the vertical dotted lines show the posterior mean of the naive regression coefficient estimates. A careful investigation of the posterior means as represented by the vertical lines on the kernel densities reveals that the adjusted for bias regression coefficient estimates are generally higher (in absolute value) than their corresponding naive estimates.

With the naive method, the variance of the regression coefficient for vegetable intake is more underestimated than for fruit intake, as depicted by the smaller length between the tails of the density plots. Of the three exposures considered in this study, the regression coefficient variance for the average number of cigarettes smoked is the most underestimated (see Table 1 and Figure 1– Figure 3). In general, a comparison of the regression coefficients' variance in the naive and the trivariate method shows that the naive method underestimates the variance of regression coefficients.

Presented in Table 2 are the mean (standard deviation) and the median for the estimates of fruit, vegetable, and the average number of cigarettes smoked adjusted for measurement error using the trivariate method in exploring

Figure 3. Kernel densities for the distribution of adjusted for measurement error and unadjusted estimates for cigarette smoking.

Table 2. The Mean (Standard Deviation) and the Median for the estimates of fruit (g/day), vegetable(g/day) and average number of cigarettes smoked per day adjusted for measurement error using the trivariate method in the sensitivity analysis by equating the limits of literature reported validity coefficients to different CIs.

CI (%)	Estimate for fruit intake		Estimate for vegetable intake	
	Mean (SD)	Median	Mean (SD)	Median
85	0.027 (0.038)	0.027	0.032 (0.051)	0.032
90	0.026 (0.037)	0.027	0.032 (0.051)	0.032
95	0.026 (0.036)	0.026	0.033 (0.051)	0.033
99	0.025 (0.035)	0.025	0.033 (0.052)	0.033
CI (%)	Estimate for smoking			
	Mean (SD)	Median		
85	-0.695 (1.839)	-0.676		
90	-0.704 (1.856)	-0.685		
95	-0.714 (1.875)	-0.695		
99	-0.727 (1.899)	-0.708		

the effects of the magnitude of uncertainty in the reported validity coefficients. From the results, the CI assumed in the distribution of the validity coefficient does not affect the mean and the median estimates of fruit, vegetable, and smoking. With the trivariate method, the results further show that the estimates' uncertainty is slightly affected by the level of uncertainty assumed for the validity coefficients. Figure 4 to Figure 6 presents the mean coefficient estimates of fruit, vegetable and the average number of cigarettes smoked adjusted for measurement error using the trivariate method in the sensitivity analysis by varying the magnitude of error correlation between measurements for the exposures (see Tables D1 to D3 in the extended data for more details²⁸).

The graphs show that varying the magnitude of the correlation between errors in any two exposures affects the estimates for the three exposures. For instance, from Figure 4, increasing the magnitude of the positive correlation

between errors in fruit and vegetable intakes increase the mean estimates for both fruit and vegetable intake while it causes a decrease (in absolute value) in the estimate for the average number of cigarettes smoked; decreasing the negative correlation between errors in the measurements for fruit and cigarette smoking decreases (in absolute value) the mean estimates for both fruit and the average number of cigarettes smoked while it leads to an increase in the estimate for vegetable intake (Figure 5). Similarly, a decrease in the magnitude of the negative correlation between errors in vegetable and number of cigarettes smoked causes a decrease (in absolute value) in the estimates for both vegetables and the average number of cigarettes smoked and an increase in the estimates for fruit intake (Figure 6).

Discussion and conclusion

In this study, we proposed and illustrated a method that adjusts for measurement error in multiple exposures measured with correlated errors in the absence of internal validation data. The method combines external validation data from the literature with the observed self-reported data to adjust for bias in the association between the

Figure 4. The mean estimates for fruit (g/day), vegetable(g/day) and average number of cigarettes smoked per day adjusted for measurement error using the trivariate method in the sensitivity analysis by varying the magnitude of error correlation between measurements for fruit and vegetable.

Figure 5. The mean estimates for fruit (g/day), vegetable(g/day) and average number of cigarettes smoked per day adjusted for measurement error using the trivariate method in the sensitivity analysis by varying the magnitude of error correlation between measurements for fruit and average number of cigarettes smoked.

Figure 6. The mean estimates for fruit (g/day), vegetable(g/day) and average number of cigarettes smoked per day adjusted for measurement error using the trivariate method in the sensitivity analysis by varying the magnitude of error correlation between measurements for vegetable and average number of cigarettes smoked.

exposures and the outcome and conduct a sensitivity analysis on the measurement error and correlation between the errors. The advantages of the multivariate method presented in this work includes: (1) the method can be used to adjust for bias in the outcome-exposure association caused by measurement error reported in multiple exposures measured with correlated errors, (2) the method is useful in the absence of the costly internal validation data, provided that external information on the correlation between the observed and the true data or the error correlations of the observed data are plausible within the study context, (3) it can be used in the sensitivity analysis on the effect of uncertainty of the reported validity coefficients, (4) can be used for sensitivity analysis on the magnitude and the direction of correlated errors, (5) the method can adjust for confounding effect in the outcome regression model and (6) This method can be easily implemented on the readily available and free software R as shown in the extended data²⁸. Often, fruit and vegetable intakes are considered as one food group. Our study is relevant because fruit intake and vegetable intake are separately assessed as independent food groups and adjusted for correlated measurement errors.

In the HBCT study example used for illustration, the estimates for fruit intake, vegetable intake, and the average number of cigarettes smoked adjusted for bias using the trivariate method were almost similar to the estimates adjusted for bias using the univariate method. The slight differences between the bias-adjusted coefficient estimates in the univariate and trivariate methods could be attributed to the weak correlations between errors assumed in this study. Sensitivity analysis on the magnitude of error correlation showed that the estimates obtained using the two methods would be different when stronger error correlations are assumed. Further, from the sensitivity analysis, we found that in a case where multiple exposures are measured with correlated errors, an increase in the magnitude of error correlation between two exposures can increase their estimates and decrease the estimate of the other exposure. From the sensitivity analysis of the level of uncertainty using CI assumed for the validity coefficients, we found that the assumed CI minimally influenced the exposures' estimates. However, the CIs for the validity coefficients should be reasonably chosen as studies have shown that uncertainty in the estimates may be affected by the level of uncertainty assigned to the validity coefficients¹⁴. From our results, we also noted that the presence of measurement error in multiple exposures can bias the association in either direction.

This study has a few limitations: (1) for simplicity, we assumed that the exposures are measured without systematic bias, i.e., only with random errors. However, in practice, the exposures can be measured with systematic error. In such a case, the systematic error components can be incorporated in the measurement error model and also in estimating the attenuation-contamination matrix; (2) although we can have a multiplicative measurement error structure⁴², our study assumed an additive measurement error structure. Exposures measured with

multiplicative error can be handled using our method by first converting the multiplicative structure to an additive structure through a suitable transformation that linearizes the error structure and (3) our study focused on a subset of current daily smokers, which is not a representative of the HBCT cohort and, therefore, the results are not generalizable.

From the findings of this study, we conclude that the multivariate method can be used to adjust for bias in the outcome-exposure association in a case where two or more exposures are measured with correlated errors. This is possible even in the absence of internal validation data provided that there is prior information about the validity of the data collection instruments and the magnitude of the measurement error correlation between the exposures. The method is useful in conducting a sensitivity analysis on the magnitude of measurement error and the sign of the error correlation.

Data availability

Source data

Data used in this study are made available to the researcher upon registration and agreeing to the terms and conditions of use in the HSRC web site at <http://curation.hsrc.ac.za/Dataset-565-datafiles.phtml>.

Extended data

Figshare: A Method to Adjust for Measurement Error in Multiple Exposures Measured with Correlated Error in the Absence of Internal Validation Study-Supplementary materials. <https://doi.org/10.6084/m9.figshare.13147970.v2>²⁸

The file shows the validity coefficient derivation, Proof for the estimate of error variance, R code for implementing the methods and convergence diagnostics results (i.e. Trace plots and ACF plots for the standard deviation and naive regression coefficient estimates of the fruits, vegetables and average number of cigarettes smoked, with explanation) and the sensitivity analysis results (supporting Tables) for varying the magnitude of error correlation between the exposures.

The extended data are available under the terms of the [Creative Commons Zero \(CC0\)](#) license.

Acknowledgements

We thank the University of KwaZulu-Natal for providing the resources needed to conduct our research. Finally, we are grateful to HSRC for allowing us to make use of their data for illustration purpose.

References

1. Kaaks R, Slimani N, Riboli E: **Pilot phase studies on the accuracy of dietary intake measurements in the epic project: overall evaluation of results. european prospective investigation into cancer and nutrition.** *Int J Epidemiol.* 1997; **26** Suppl 1: S26–36. [PubMed Abstract](#) | [Publisher Full Text](#)
2. Collese T, Vatauvuk-Serrati G, Nascimento-Ferreira M, et al.: **What is the validity of questionnaires assessing fruit and vegetable consumption in children when compared with blood biomarkers? a meta-analysis.** *Nutrients.* 2018; **10**(10): 1396. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
3. Goldbohm RA, van den Brandt PA, Brants HAM, et al.: **Validation of a dietary questionnaire used in a large-scale prospective cohort study on diet and cancer.** *Eur J Clin Nutr.* 1994; **48**(4): 253–265. [PubMed Abstract](#)
4. Plaete J, De Bourdeaudhuij I, Crombez G, et al.: **The reliability and validity of short online questionnaires to measure fruit and vegetable intake in adults: the fruit test and vegetable test.** *PLoS One.* 2016; **11**(7): e0159834. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
5. Agudo A: **Measuring intake of fruit and vegetables.** 2005; Accessed Mar.10,2020. [Reference Source](#)
6. Kipnis V, Subar AF, Midthune D, et al.: **Structure of dietary measurement error: results of the open biomarker study.** *Am J Epidemiol.* 2003; **158**(1): 14–21; discussion 22–6. [PubMed Abstract](#) | [Publisher Full Text](#)
7. Gleser LJ: **Improvements of the naive approach to estimation in nonlinear errors-in-variables regression models.** *Contemp Math.* 1990; **112**: 99–114. [Reference Source](#)
8. Day NE, McKeown N, Wong MY, et al.: **Epidemiological assessment of diet: a comparison of a 7-day diary with a food frequency questionnaire using urinary markers of nitrogen, potassium and sodium.** *Int J Epidemiol.* 2001; **30**(2): 309–317. [PubMed Abstract](#) | [Publisher Full Text](#)
9. Subar AF, Kipnis V, Troiano RP, et al.: **Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: the open study.** *Am J Epidemiol.* 2003; **158**(1): 1–13. [PubMed Abstract](#) | [Publisher Full Text](#)
10. Natarajan L, Pu M, Fan J, et al.: **Measurement error of dietary self-report in intervention trials.** *Am J Epidemiol.* 2010; **172**(7): 819–827. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
11. Kipnis V, Freedman LS, Carroll RJ, et al.: **A bivariate measurement error model for semicontinuous and continuous variables: Application to nutritional epidemiology.** *Biometrics.* 2016; **72**(1): 106–115. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
12. Carroll RJ, Ruppert D, Stefanski LA, et al.: **Measurement error**

- in nonlinear models: a modern perspective. CRC press, 2006.
[Reference Source](#)
13. Kaaks R, Riboli E, van Staveren W: **Calibration of dietary intake measurements in prospective cohort studies.** *Am J Epidemiol.* 1995; **142**(5): 548–556.
[PubMed Abstract](#) | [Publisher Full Text](#)
 14. Agogo GO, van der Voet H, van't Veer P, et al.: **A method for sensitivity analysis to assess the effects of measurement error in multiple exposure variables using external validation data.** *BMC Med Res Methodol.* 2016; **16**(1): 139.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 15. Dellaportas P, Stephens DA: **Bayesian analysis of errors-in-variables regression models.** *Biometrics.* 1995; **51**(3): 1085–1095.
[Publisher Full Text](#)
 16. Huang Y, Chen J, Qiu H: **Bayesian quantile regression for nonlinear mixed-effects joint models for longitudinal data in the presence of mismeasured covariate errors.** *J Biopharm Stat.* 2017; **27**(5): 741–755.
[PubMed Abstract](#) | [Publisher Full Text](#)
 17. Lin X: **A Bayesian semiparametric accelerated failure time model for arbitrarily censored data with covariates subject to measurement error.** *Commun Stat Simul Comput.* 2017; **46**(1): 747–756.
[Publisher Full Text](#)
 18. Muff S, Ott M, Braun J, et al.: **Bayesian two-component measurement error modelling for survival analysis using inla—a case study on cardiovascular disease mortality in Switzerland.** *Comput Stat Data An.* 2017; **113**: 177–193.
[Publisher Full Text](#)
 19. Rosner B, Willett WC, Spiegelman D: **Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error.** *Stat Med.* 1989; **8**(9): 1051–1069; discussion 1071–3
[PubMed Abstract](#) | [Publisher Full Text](#)
 20. R Core Team: **R: A Language and Environment for Statistical Computing.** R Foundation for Statistical Computing, Vienna, Austria. 2018; **2**.
[Reference Source](#)
 21. Van Heerden AC: **Non-communicable disease screening and HIV testing and counselling in rural KwaZulu-Natal, South Africa (NCD) 2015.** [Data set]. NCD 2015. Version 1.0.: Human Sciences Research Council [distributor]. 2016.
[Publisher Full Text](#)
 22. Barnabas RV, van Rooyen H, Tumwesigye E, et al.: **Initiation of antiretroviral therapy and viral suppression after home hiv testing and counselling in kwazulu-natal, south africa, and mbarara district, uganda: a prospective, observational intervention study.** *Lancet HIV.* 2014; **1**(2): e68–e76.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 23. Newby PK, Muller D, Hallfrisch J, et al.: **Dietary patterns and changes in body mass index and waist circumference in adults.** *Am J Clin Nutr.* 2003; **77**(6): 1417–1425.
[PubMed Abstract](#) | [Publisher Full Text](#)
 24. Field AE, Gillman MW, Rosner B, et al.: **Association between fruit and vegetable intake and change in body mass index among a large sample of children and adolescents in the united states.** *Int J Obes Relat Metab Disord.* 2003; **27**(7): 821–826.
[PubMed Abstract](#) | [Publisher Full Text](#)
 25. Azagba S, Sharaf MF: **Fruit and vegetable consumption and body mass index: a quantile regression approach.** *J Prim Care community Health.* 2012; **3**(3): 210–220.
[PubMed Abstract](#) | [Publisher Full Text](#)
 26. Yirga AA, Ayele DG, Melesse SF: **Application of quantile regression: Modeling body mass index in ethiopia.** *The Open Public Health Journal.* 2018; **11**(1): 221–233.
[Publisher Full Text](#)
 27. Freedman LS, Schatzkin A, Midthune D, et al.: **Dealing with dietary measurement error in nutritional cohort studies.** *J Natl Cancer Inst.* 2011; **103**(14): 1086–1092.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 28. Muoka AK, Agogo GO, Ngesa OO, et al.: **A method to adjust for measurement error in multiple exposures measured with correlated error in the absence of internal validation study-supplementary materials.** 2020.
<http://www.doi.org/10.6084/m9.figshare.13147970.v2>
 29. Hoff PD: **A first course in Bayesian statistical methods.** Springer, 2009; **580**.
[Publisher Full Text](#)
 30. Plummer M: **rjags: Bayesian Graphical Models using MCMC.** R package version 4-8. 2018.
[Reference Source](#)
 31. Lunn D, Spiegelhalter D, Thomas A, et al.: **The bugs project: Evolution, critique and future directions.** *Stat Med.* 2009; **28**(25): 3049–3067.
[PubMed Abstract](#) | [Publisher Full Text](#)
 32. Martin AD, Quinn KM, Park JH: **MCMCpack: Markov chain monte carlo in R.** *J Stat Softw.* 2011; **42**(9): 22.
[Publisher Full Text](#)
 33. Plummer M, Best N, Cowles K, et al.: **Coda: convergence diagnosis and output analysis for mcmc.** *R news.* 2006; **6**(1): 7–11.
[Reference Source](#)
 34. Plummer M, Stukalov A, Denwood M, et al.: **Package 'rjags'.** *update.* 16: 1, 2018.
 35. Feskanih D, Rimm EB, Giovannucci EL, et al.: **Reproducibility and validity of food intake measurements from a semiquantitative food frequency questionnaire.** *J Am Diet Assoc.* 1993; **93**(7): 790–796.
[PubMed Abstract](#) | [Publisher Full Text](#)
 36. Woodward M, Moohan M, Tunstall-Pedoe H: **Self-reported smoking, cigarette yields and inhalation biochemistry related to the incidence of coronary heart disease: results from the scottish heart health study.** *J Epidemiol Biostat.* 1999; **4**(4): 285–295.
[PubMed Abstract](#)
 37. Eliopoulos C, Klein J, Koren G: **Validation of self-reported smoking by analysis of hair for nicotine and cotinine.** *Ther Drug Monit.* 1996; **18**(5): 532–536.
[PubMed Abstract](#) | [Publisher Full Text](#)
 38. Secker-Walker RH, Vacek PM, Flynn BS, et al.: **Exhaled carbon monoxide and urinary cotinine as measures of smoking in pregnancy.** *Addict Behav.* 1997; **22**(5): 671–684.
[PubMed Abstract](#) | [Publisher Full Text](#)
 39. Stram DO, Huberman M, Wu AH: **Is residual confounding a reasonable explanation for the apparent protective effects of beta-carotene found in epidemiologic studies of lung cancer in smokers?** *Am J Epidemiol.* 2002; **155**(7): 622–628.
[PubMed Abstract](#) | [Publisher Full Text](#)
 40. Fisher RA: **Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population.** *Biometrika.* 1915; **10**(4): 507–521.
[Publisher Full Text](#)
 41. Fisher RA: **On the 'probable error' of a coefficient of correlation deduced from a small sample.** *Metron.* 1921; **1**: 1–32.
[Reference Source](#)
 42. Heid IM, Küchenhoff H, Miles J, et al.: **Two dimensions of measurement error: classical and berkson error in residential radon exposure assessment.** *J Expo Anal Environ Epidemiol.* 2004; **14**(5): 365–77.
[PubMed Abstract](#) | [Publisher Full Text](#)

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 25 March 2022

<https://doi.org/10.5256/f1000research.30843.r121761>

© 2022 Molebatsi K. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Kesaobaka Molebatsi

University of Botswana, Notwane Rd, Gaborone, Botswana

The authors have proposed a method that accounts for measurement errors in multiple exposures that are correlated and called it a multivariate method. They have clarified the challenges of ignoring such a problem well in the absence of validation samples, and have compared the multivariate method with other available methods. Their findings from a real data set suggest that the multivariate method can be used to adjust for the bias in the absence of an internal validation study.

Both advantages and disadvantages of the multivariate method have been clearly discussed. However, I expected the authors to conduct a simulation study to validate their method further. This is because the real data set falls short to give appropriate information regarding model performance as the true population parameters are unknown. Moreover, it is just one of the many possible samples that one can get from the same population. I am not insisting that the authors should conduct these extra and possibly time-consuming analyses, but they can somehow mention the lack of it as a potential limitation.

Otherwise, the paper is well written and contributes towards solving an important statistical problem with a sound method.

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Biostatistical methods, particularly of dealing with confounding, selection bias, measurement or misclassification error and interference.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 06 May 2021

<https://doi.org/10.5256/f1000research.30843.r84307>

© 2021 Ponzi E. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Erica Ponzi

Oslo Center for Biostatistics and Epidemiology, University of Oslo, Oslo, Norway

The paper presents an application of measurement error modeling to a dataset from a home-based HIV counseling and testing (HBCT) study. It focuses on the case where multiple variables are measured with error and such errors are correlated.

The proposed model is not novel in itself, as Bayesian measurement error models have been used before and extended to the case of multivariate cases, but the application is interesting and the use of the Fisher transformation for the correlation among errors hasn't been employed in these specific cases before.

I believe this can be an interesting contribution to the field, and the implementation of the presented model can be used in similar studies, which are becoming very common in epidemiology. Nevertheless, since the paper is presented as a method paper, I think more methodological aspects should be examined:

1. Is each error-prone variable assumed to have a classical measurement error structure? This is not explicitly said in the paper, but I think this aspect deserves more attention. Is it reasonable to assume a classical measurement error for all the three variables? Wouldn't a Berkson error, or a mixture of the two also make sense? If we think about some kind of "rounding" error, which can be plausible in these cases, a Berkson structure would seem appropriate. It is known that in the presence of a single variable measured with additive Berkson error, and uncorrelated to other

variables and to the response, the attenuation problem does not occur but only an increase of uncertainty is observed. In the case of multiple, correlated, errors this is not obvious so I believe such situation should also be explored, and similar models with a Berkson or a mixture error structure should be investigated (or at least the attenuation phenomenon in such cases).

2. Not all measurement error techniques correct for attenuation simply by dividing by the attenuation factor, see for example the simulation extrapolation technique or the hierarchical Bayesian measurement error models. Adding a latent level for the error eg in a Bayesian framework does not require the attenuation factor to be modeled explicitly and allows for different error structures (see point 1 above) and different correlation structures. The proposed model explicitly estimates the attenuation factor but a more general latent error model can be incorporated to allow for broader applicability (see Bayesian measurement error models in the Carroll book or in some of the cited papers). The author should explain their choice in more details, and try to accommodate different error structures in their model.

3. The response variable is BMI, ie a continuous variable. What would happen if the response was binary? Do the proposed models generalize to such case?

3. What would happen if error is correlated with the response? Worth discussing this aspect even if models do not explicitly have to account for it.

4. It seems that smoking has a higher effect on BMI than vegetables and fruit. Is this supported by other findings?

5. Have you tried with a higher thinning interval? It avoids correlation between samples.

6. It would be interesting to see some sensitivity analysis on the priors too.

7. In Figure 4,5,6 the x-axis correlation between errors. It's difficult to see the effect because of the different scales of smoking and fruit/vegetables. Maybe it would be clearer to have one plot per beta and three lines in each plot corresponding to the error correlation.

8. I agree with the authors that modeling measurement error brings additional uncertainty in the parameter estimation. This trade-off between bias and variance is a commonly discussed aspect of measurement error modeling. Bayesian models allow to directly incorporate the uncertainty about the error in the posterior estimates of interest. I believe this aspect should be elaborated in the discussion.

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Partly

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Biostatistics, measurement error, randomized clinical trials

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research