A Method to Adjust for Measurement Error in Three Exposures Measured with Correlated Errors in the Absence of Internal Validation Study

Alexander K. Muoka (alexanderkasyoki@ttu.ac.ke)
University of KwaZulu-Natal School of Mathematics, Statistics and Computer Science
https://orcid.org/0000-0002-8768-1413

George Agogo
Yale University

Oscar Ngesa
Taita Taveta University

Henry Mwambi
University of KwaZulu-Natal School of Mathematics, Statistics and Computer Science

Research article

Keywords: Measurement error, Internal validation study, Attenuation, Bias, Questionnaire data

Posted Date: November 11th, 2019

DOI: https://doi.org/10.21203/rs.2.16959/v1

RESEARCH

A Method to Adjust for Measurement Error in Three Exposures Measured with Correlated Errors in the Absence of Internal Validation Study

Alexander K. Muoka[1,3*†], George O. Agogo[2], Oscar Ngesa[3] and Henry Mwambi[1]

*Correspondence: alexanderkasyoki@gmail.com
[1]School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, 3209 Pietermaritzburg, South Africa
Full list of author information is available at the end of the article
†This article constitutes part of the corresponding author's Ph.D. thesis at the University of KwaZulu-Natal, South Africa.

Abstract

Difficulty in obtaining the correct measurement for an individual's long-term 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 proposed a method (trivariate method) that adjusts for measurement error in three correlated exposures in the absence of internal validation study and illustrated the method using real data. We compared the results from the proposed method with those obtained using a method that ignores measurement error and a method that ignores correlations between the errors and true exposures (the univariate method). It was found that ignoring measurement error leads to bias and underestimates the standard error. It was also found that the magnitude of adjustment in the trivariate method is sensitive to the magnitude of measurement error, sign and correlation between the errors. We conclude that the proposed method can be used to adjust for bias in the outcome-exposure association in a case where three exposures are measured with correlated errors in the absence of an internal validation study. The method is 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

Background

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 an 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 vegetables [1, 2, 3, 4, 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 questionnaire [7, 8, 9, 6, 10, 11]. Validation study constitutes a small number of individuals from whom dietary intakes are measured repeatedly using unbiased instrument [12]. However, these studies are expensive to conduct and, in some cases not feasible. Several methods have been proposed to handle measurement error in the absence of internal validation data [13, 14, 15, 16, 17].

Agogo et al. [13] conducted a sensitivity analysis to investigate the effect of the magnitude of 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 Stephens [14] presented a Bayesian method for analysis of non-linear errors-in-variable where prior knowledge of the unknown true covariate is incorporated. Huang et al. [15] proposed a quantile regression-based non-linear mixed effects joint models for longitudinal data that simultaneously accounts for response with non-central location and for covariate with non-normality and measurement error under Bayesian framework. Lin [16] 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. [17] introduced a Bayesian method to handle a mixture of classical and Berkson measurement errors in a single explanatory variable and illustrated their method in studying cardiovascular disease mortality.

Majority of these authors considered a case where one exposure is measured with error (hereafter, a univariate case); Agogo et al. [13] focused on a case where two exposures are measured with error (hereafter, a bivariate case). In a univariate case, the bias in the association between an outcome and the exposure is adjusted by dividing the unadjusted association estimate by the attenuation factor [18]. 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 extend bias adjustment methods to a case where three exposures are measured with correlated errors (hereafter, a trivariate case) in the absence of internal validation study and demonstrate the implementation of this method using `R` software [19]. In this case, we use real data to illustrate the method. Specifically, we use a subset data from a home-based HIV counseling and testing study that was conducted in rural and peri-urban communities in KwaZulu-Natal Province, South Africa [20]. Unlike the methods proposed in the literature, we developed an `R` package for our method.

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.

Materials and 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 2012 [20]. The data were obtained from the Human Sciences Research Council (HSRC) of South Africa [20]. This study aimed 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 world [20].

Home-based HIV counseling and testing is a cross-sectional, single site study in South Africa which aims to increase engagement in HIV care by integrating NCDs screening with community-based HIV testing [21]. A random sampling approach was used, where 587 participants over the age of 18 were selected from 50,000 people living in Mpumuza suburb [20]. 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 Mpumuza 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 analysis [20].

In our study, we used a subset data consisting of 76 current daily-smokers of cigarette to model the amount of association between body mass index (BMI) and three exposures namely: smoking, fruit and vegetable consumption. BMI was measured in kg/m^2, while smoking was measured as the average number of cigarettes smoked per day. Initially, fruits and vegetable intake were measured in number of servings consumed per day. It is often assumed that a standard portion of fruit/vegetable weighs about 80g [5]. 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. In this set up there are two draw backs: (1) measurement error in the recorded number of cigarettes smoked due to possible misreporting and (2) measurement error in fruits and vegetable consumption due to recall bias and assuming the average weight of a portion of fruit or vegetable. The subset data used in this study is not a representative of the the HBCT-NCD cohort and is only used to illustrate how to adjust for correlated measurement error in the absence of internal validation data and not for inferential purpose.

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 background and objectives of the study) and their rights regarding participation and withdrawing 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, defined using the following generalized linear model

$$g[E(Y|X_1, X_2, X_3)] = \beta_0 + \beta_{X_1} X_1 + \beta_{X_2} X_2 + \beta_{X_3} X_3, \quad (1)$$

where g(.) is the identity link function, Y denotes the BMI, β_0 is the 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. 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 $\hat{\boldsymbol{\beta}_W} = (\hat{\beta}_{W_1}, \hat{\beta}_{W_2}, \hat{\beta}_{W_3})^\top$.

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

$$\boldsymbol{W}_i = \boldsymbol{\alpha_0}_i + \boldsymbol{\alpha_1}_i X_i + \boldsymbol{\epsilon_W}_i,\ i = 1, 2, 3 \tag{2}$$

where $\boldsymbol{\epsilon_W} = (\epsilon_{W_1}, \epsilon_{W_2}, \epsilon_{W_3})^\top$, $\boldsymbol{\epsilon_W} \sim \mathrm{N}(\mathbf{0}, \Sigma_{\boldsymbol{\epsilon_W}})$; $\mathbf{W} = (W_1, W_2, W_3)^\top$; $\boldsymbol{\alpha}_0 = (\alpha_{01}, \alpha_{02}, \alpha_{03})^\top$, $\boldsymbol{\alpha}_1 = (\alpha_{11}, \alpha_{12}, \alpha_{13})^\top$; with the terms in $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_1$ quantifying the constant bias and the proportional scaling bias respectively; $\boldsymbol{\epsilon_W}$ 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, bias in the association between an outcome and an exposure is adjusted by dividing the unadjusted association estimate by the attenuation factor [18]. Attenuation factor (λ) is defined as, $\lambda = var(X_i)/var(W_i)$, 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.

A bivariate method

In the case where two exposures are measured with correlated errors (hereafter, a bivariate case), bias in the association between an outcome and the exposures can be adjusted using the relationship, $\beta_{\boldsymbol{W}2} = (\Lambda_2^T)^{-1}\beta_{\boldsymbol{X}}$, where Λ_2 denotes a 2×2 attenuation-contamination matrix [18, 22]. The off-diagonal elements in Λ are known as contamination factors while the diagonal elements are called attenuation factors [13]. Therefore, the bivariate method considers the contamination effect caused by correlated measurement errors. Noteworthy, this method ignores measurement error and error correlation for the third exposure variable, when dealing with three exposures measured with error.

The proposed method

For simplicity and without loss of generality, we assume that W_i is measured without systematic bias (i.e., $\alpha_{0i} = 0$, $\alpha_{1i} = 1$ for the three exposures). When dealing

with three exposures measured with correlated errors (hereafter, trivariate case), we define the estimate of attenuation-contamination matrix $\hat{\Lambda}$ as

$$\hat{\Lambda} = \underbrace{\begin{bmatrix} \hat{\sigma}^2_{X_1} & \hat{\sigma}_{X_1X_2} & \hat{\sigma}_{X_1X_3} \\ \hat{\sigma}_{X_1X_2} & \hat{\sigma}^2_{X_2} & \hat{\sigma}_{X_2X_3} \\ \hat{\sigma}_{X_1X_3} & \hat{\sigma}_{X_2X_3} & \hat{\sigma}^2_{X_3} \end{bmatrix}}_{\widehat{\Sigma}_{\boldsymbol{X}}} \underbrace{\begin{bmatrix} \hat{\sigma}^2_{W_1} & \hat{\sigma}_{W_1W_2} & \hat{\sigma}_{W_1W_3} \\ \hat{\sigma}_{W_1W_2} & \hat{\sigma}^2_{W_2} & \hat{\sigma}_{W_2W_3} \\ \hat{\sigma}_{W_1W_3} & \hat{\sigma}_{W_2W_3} & \hat{\sigma}^2_{W_3} \end{bmatrix}^{-1}}_{\widehat{\Sigma}^{-1}_{\boldsymbol{W}}}, \tag{3}$$

where $\widehat{\Sigma}_{\boldsymbol{X}}$ is the estimate of covariance matrix of the true intakes, $\widehat{\Sigma}^{-1}_{\boldsymbol{W}}$ is the inverse of the estimate of covariance matrix of the measured exposures, $\hat{\sigma}^2_{X_i}$ is the variance estimate of X_i $(i = 1, 2, 3)$; $\hat{\sigma}_{X_iX_j}$ $(i \neq j)$ denotes the covariance estimate between the true exposures; $\hat{\sigma}^2_{W_i}$ is the variance estimate of W_i ; $\hat{\sigma}_{W_iW_j}$ $(i \neq j)$ is the covariance estimate between the observed exposures.

In the trivariate case, we can obtain the adjusted association estimates by pre-multiplying the unadjusted association estimates by the inverse of the transpose of attenuation-contamination matrix as

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{X}} = (\hat{\Lambda}^T)^{-1}\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}, \tag{4}$$

where $\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}$ can be obtained from the observed questionnaire data.

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

$$\begin{aligned} \rho_{W_iX_i} &= \frac{cov(W_i,\ X_i)}{\sqrt{var(W_i)var(X_i)}}, \quad i = 1, 2, 3. \\ &= \frac{\sigma_{X_i}}{\sigma_{W_i}}, \end{aligned} \tag{5}$$

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

$$\hat{\sigma}^2_{X_i} = \left(\hat{\rho}_{W_iX_i}\hat{\sigma}_{W_i}\right)^2, \tag{6}$$

by incorporating external validation information on $\rho_{W_iX_i}$.

To obtain covariances between the true exposures (i.e. $\hat{\sigma}_{X_1X_2}$, $\hat{\sigma}_{X_1X_3}$ and $\hat{\sigma}_{X_2X_3}$), one of the following two approaches is used: (i) if external information about $\hat{\rho}_{X_iX_j}$ is available, we obtain covariances between true exposures as follows:

$$\begin{aligned} \hat{\sigma}_{X_1X_2} &= \hat{\rho}_{X_1X_2}\hat{\sigma}_{X_1}\hat{\sigma}_{X_2} \\ \hat{\sigma}_{X_1X_3} &= \hat{\rho}_{X_1X_3}\hat{\sigma}_{X_1}\hat{\sigma}_{X_3} \\ \hat{\sigma}_{X_2X_3} &= \hat{\rho}_{X_2X_3}\hat{\sigma}_{X_2}\hat{\sigma}_{X_3}, \end{aligned} \tag{7}$$

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

$$\begin{aligned}\hat{\sigma}_{W_iW_j} &= \hat{\sigma}_{X_iX_j} + \hat{\sigma}_{\epsilon_{W_i}\epsilon_{W_j}} + \underbrace{\sigma_{X_i\epsilon_{W_j}}}_{0} + \underbrace{\sigma_{X_j\epsilon_{W_i}}}_{0} \\ &= \hat{\sigma}_{X_iX_j} + \hat{\rho}_{\epsilon_{W_i}\epsilon_{W_j}}\hat{\sigma}_{\epsilon_{W_i}}\hat{\sigma}_{\epsilon_{W_j}}, \end{aligned} \tag{8}$$

where X_i and ϵ_{W_j}, X_j and ϵ_{W_i} are assumed to be uncorrelated.
From equations (2) and (6), the estimate of the error variance $\hat{\sigma}^2_{\epsilon_{W_i}}$ is

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

See Additional file 1: Appendix B for the proof.
From equations (8-9), the covariances between the true exposures are given by

$$\begin{aligned}\hat{\sigma}_{X_1X_2} =& \hat{\sigma}_{W_1W_2} - \hat{\rho}_{\epsilon_{W_1}\epsilon_{W_2}}\hat{\sigma}_{W_1}\hat{\sigma}_{W_2}\sqrt{(1-\hat{\rho}^2_{W_1X_1})(1-\hat{\rho}^2_{W_2X_2})} \\ \hat{\sigma}_{X_1X_3} =& \hat{\sigma}_{W_1W_3} - \hat{\rho}_{\epsilon_{W_1}\epsilon_{W_3}}\hat{\sigma}_{W_1}\hat{\sigma}_{W_3}\sqrt{(1-\hat{\rho}^2_{W_1X_1})(1-\hat{\rho}^2_{W_3X_3})} \\ \hat{\sigma}_{X_2X_3} =& \hat{\sigma}_{W_2W_3} - \hat{\rho}_{\epsilon_{W_2}\epsilon_{W_3}}\hat{\sigma}_{W_2}\hat{\sigma}_{W_3}\sqrt{(1-\hat{\rho}^2_{W_2X_2})(1-\hat{\rho}^2_{W_3X_3})}. \end{aligned} \tag{10}$$

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 (3) and adjust for the bias in the association between the exposures measured with error and the outcome using equation (4).

Illustration of the proposed method using the study data

We illustrate the 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 (hereafter, trivariate method). MCMC is a class of algorithms that samples from the posterior distributions by traversing the parameter space [23]. Posterior distribution is obtained by updating the prior distribution with observed data. The steps for implementing the proposed 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, we estimated the posterior distribution of the covariance matrix for the observed exposures ($\Sigma_{\boldsymbol{W}}$). The exposures were assumed to follow a multivariate normal distribution with mean and covariance, i.e. $\boldsymbol{W} \sim \mathrm{N}_3(\mu_{\boldsymbol{W}}, \Sigma_{\boldsymbol{W}})$. We assumed weakly informative multivariate normal prior for $\mu_{\boldsymbol{W}}$ as $\mu_{\boldsymbol{W}_{prior}} \sim \mathrm{N}_3(0, 10^6\boldsymbol{I}_3)$, where $\boldsymbol{I}_3$ is a 3×3 identity matrix. In multivariate normal distribution, $\Sigma_{\boldsymbol{W}}$ must satisfy two conditions: (1) be positive definite (i.e. $\boldsymbol{W}^T\Sigma_{\boldsymbol{W}}\boldsymbol{W} > 0$, for all $\boldsymbol{W}$) and (2) be a symmetric matrix. The semi-conjugate prior distribution for $\Sigma_{\boldsymbol{W}}$, which has these two properties is the inverse-wishart distribution [23]. To minimize influence of the prior information on the estimate of $\Sigma_{\boldsymbol{W}}$, we considered weakly informative inverse-wishart prior as $\Sigma_{\boldsymbol{W}_{prior}} \sim \mathrm{IW}(\boldsymbol{I}_3, v)$, where $v = 3$ 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 distribution of the variance of true intakes ($\hat{\sigma}^2_{X_i}$) using equation (5). To estimate the covariance between true intakes ($\hat{\sigma}_{X_iX_j}$) using equation (9), we required external validation information on correlation between the errors ($\rho_{\epsilon_{W_i}\epsilon_{W_j}}$). Similar to Agogo et al.[13], we generated the correlation between errors from 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) from their joint distribution as shown in equation (3).

Lastly, we fitted a Bayesian multiple linear regression model (hereafter, naive method) to obtain the posterior distributions of the unadjusted coefficient estimates $(\hat{\beta}_{W_1}, \hat{\beta}_{W_2}, \hat{\beta}_{W_3})^\top$. 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 $\beta_{W_i\ prior} \sim \mathrm{N}(0, 10^6)$. The adjusted coefficient estimates $\hat{\boldsymbol{\beta}}_{\boldsymbol{X}}$ were then obtained from the joint posterior distribution of $\hat{\Lambda}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}$ as $\hat{\boldsymbol{\beta}}_{\boldsymbol{X}} = (\hat{\Lambda}^T)^{-1}\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}$.

Software implementation of the proposed method

We implemented the proposed method in `R` using `rjags`, `coda`, `MCMCpack` and `mvtnorm` packages. To facilitate Bayesian estimation of the covariance matrix of the observed exposures ($\Sigma_{\boldsymbol{W}}$), `rjags` package was used to provide an interface from R to the JAGS library [24]. JAGS is a gibbs sampler that uses MCMC to draw dependent samples from the posterior distribution of the parameters [25]. The Bayesian estimation of $\Sigma_{\boldsymbol{W}}$ proceeded in the following steps: (1) defining a model for $\Sigma_{\boldsymbol{W}}$ 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 `MCMCpack` package was used to generate a posterior density sample from the naive linear regression model [26]. MCMC convergence diagnostics of all the model parameters was done using trace plots and autocorrelation (ACF) plots from the `coda` package [27]. See Additional file 1: Appendix D for convergence diagnostics results. For each model, the burn-in iterations was 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 performance [28]. 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, median and 95% CIs. We compared the results obtained under the method that ignores measurement error, the univariate and trivariate methods. The `R code` used for analysis is presented in Additional file 1: Appendix C.

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 ranged 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, 29]. 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 women [30], the correlation between 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 cotinine [31], 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 number of cigarettes smoked and carboxy-haemoglobin 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 women [32], a validity coefficient in the range of 0.61 to 0.70 was reported. A study by Stram et al. [33] found the correlation between 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.

Similar to Agogo et al. [13], 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:

a. 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;
b. 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 tends 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 to the covariance between the true intakes [13].

Estimating the distribution of $\rho_{W_iX_i}$

Using the range of plausible values obtained from external validation information, we generated the validity coefficients using 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 [34, 35]. The procedure is as outlined below:

(i) Using the Fisher Z-transformation formula

$$F_{Z_i} = 0.5\left[\ln(1+\rho_{W_iX_i}) - \ln(1-\rho_{W_iX_i})\right], \tag{11}$$

transform the lower (r_l) and upper (r_u) limits of the validity coefficient $\rho_{W_iX_i}$ 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 $\sigma_{Z_i} = \dfrac{0.5(F_{Z_u} - F_{Z_l})}{Z_{\alpha/2}}$, where $Z_{\alpha/2}$ is the $(1-\dfrac{\alpha}{2})\%$ quantile of a standard normal random variable.

(iii) Generate F_{Z_i}'s as $F_{Z_i} \sim N(\mu_{Z_i}, \sigma^2_{Z_i})$

(iv) Using the inverse of Fisher Z-transformation, back-transform the generated F_{Z_i}'s to validity coefficient as

$$\rho_{W_iX_i} = \frac{\exp(2F_{Z_i}) - 1}{\exp(2F_{Z_i}) + 1}. \tag{12}$$

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 to assess the sensitivity of the estimates to various magnitudes of CI and correlation between errors, when using the proposed method.

Results

Table 1 presents regression coefficients estimates for fruit intake (g/day), vegetable intake (g/day) and the average number of cigarettes smoked per 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 average number of cigarettes smoked, the bias adjusted coefficient estimates were three times as large as the naive coefficient estimates. For vegetables intake, the increase in the strength of

the association was about four times as compared to the naive regression coefficient estimates.

[Table 1]

For both fruit intake and average number of cigarettes smoked, the univariate method gave slightly greater estimates while the bias adjusted values for vegetable intake was slightly lower in the univariate method. The variability for regression coefficient estimate of the number of cigarettes smoked was higher than that for both fruits and vegetables 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.

Figures 1-3 shows 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 respectively. The solid vertical lines on the density plots depicts 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.

[Figure 1]

[Figure 2]

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 variance of the regression coefficient for the average number of cigarettes smoked is the most underestimated (see Table 1 and Figures 1 - 3). In general, a comparison of the variance of the regression coefficients in the naive and the proposed method shows that the naive method underestimates the variance of regression coefficients.

[Figure 3]

Presented in Table 2 is the mean (standard deviation), median and the 95% CI for the estimates of fruit, vegetable and the average number of cigarettes smoked adjusted for measurement error using the trivariate (proposed) method in exploring the effects of the magnitude of uncertainty in the reported validity coefficients. From the results, the CI assumed in the distribution of validity coefficient does not affect the mean and the median estimates of fruit, vegetable and smoking. With the proposed method, the results further shows that the uncertainty in the estimates is slightly affected by the level of uncertainty assumed for the validity coefficients.

[Table 2]

Tables 3 to 5 presents the mean (standard deviation), median and 95% CI for the 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.

[Table 3]

[Table 4]

The results shows that varying the magnitude of correlation between errors in any two exposures affects the estimates for the three exposures. For instance, from

Table 3 increasing the magnitude of the positive correlation between errors in fruit and vegetable intake increases the mean and the median 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; increasing the negative correlation between errors in the measurements for fruit and cigarette smoking increases (in absolute value) both the mean and the median estimates for both fruit and the average number of cigarettes smoked while it leads to a decrease in the estimate for vegetable intake (Table 4). Similarly, an increase in the magnitude of the negative correlation between errors in vegetable and fruit intake causes an increase in the estimates for both vegetable and cigarette smoking and a decrease in the estimate for fruit intake (Table 5).

[Table 5]

Discussion and Conclusion

In this study, we proposed and illustrated a method that adjusts for measurement error in three exposures measured with correlated errors in the absence of internal validation data. The proposed method combines external validation data from the literature with the observed self-reported data to adjust for bias in the association between the exposures and the outcome. The advantages of the trivariate method proposed 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 three exposures and can be extended to more than three 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 shown in Additional file 1: Appendix C. 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 will be different when stronger error correlations are assumed. Further, from the sensitivity analysis, we found that in a case where three exposures are measured with correlated errors, an increase in the magnitude of error correlation between two exposures can increase their estimates and cause a decrease in the estimate of the other exposure. From the sensitivity analysis of the level of level of

uncertainty using CI assumed for the validity coefficients, we found that the estimates for the exposures were minimally influenced by the assumed CI. 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 coefficients [13]. From our results, we also noted that the presence of measurement error in the exposures can bias the association in either direction. These results are in support of the finding by Agogo et.al [13] that it is difficult to predict the direction and magnitude of the association between the exposure(s) of interest when several exposures are measured with correlated errors. Noteworthy, our results shows that both fruit and vegetable intakes have a weak positive association with BMI. This is contrary to the expected. However, it is worth noting that some types of vegetables and fruits have high sugar content in them which can be associated with slight increase in BMI if consumed in excess [36, 37].

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 structure [38], 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 proposed method can be used to adjust for bias in the outcome-exposure association in a case where three 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.

Additional File

Additional file 1

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). [PDF 484KB]

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

Funding

This work was supported through the DELTAS Africa Initiative Grant No. 107754/Z/15/Z-DELTAS Africa SSACAB. 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 (Grant No. 107754/Z/15/Z) and the UK government. The views expressed in this publication are those of the author(s) and not necessarily those of AAS, NEPAD Agency, Wellcome Trust or the UK government.

Data availability
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.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Author's contributions
AKM HM GOA ON conceived the idea. AKM contributed in developing the method, wrote the R code, analysed the data and wrote the draft manuscript. G0A, HM and ON helped in developing the method and writing the paper.

Acknowledgements
We thank in advance the editors of this work and the anonymous reviewers for their helpful comments to improve the work. 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 purposes.

Author details
[1]School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, 3209 Pietermaritzburg, South Africa. [2]Centers for Disease Control, Nairobi, Kenya. [3] School of Science and Informatics, Taita Taveta University, 635-80300 Voi, Kenya.

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. International journal of epidemiology **26**(suppl_1), 26 (1997)
2. Collese, T., Vatavuk-Serrati, G., Nascimento-Ferreira, M., De Moraes, A., Carvalho, H.: What is the validity of questionnaires assessing fruit and vegetable consumption in children when compared with blood biomarkers? a meta-analysis. nutrients **10**(10), 1396 (2018)
3. Goldbohm, R.A., van den Brandt, P.A., Brants, H.A., van't Veer, P., Al, M., Sturmans, F., Hermus, R.: Validation of a dietary questionnaire used in a large-scale prospective cohort study on diet and cancer. European journal of clinical nutrition **48**(4), 253–265 (1994)
4. Plaete, J., De Bourdeaudhuij, I., Crombez, G., Steenhuyzen, S., Dejaegere, L., Vanhauwaert, E., Verloigne, M.: The reliability and validity of short online questionnaires to measure fruit and vegetable intake in adults: the fruit test and vegetable test. PloS one **11**(7), 0159834 (2016)
5. Agudo, A.: Measuring Intake of Fruit and Vegetables. [electronic source]. https://apps.who.int/iris/handle/10665/43144
6. Kipnis, V., Subar, A.F., Midthune, D., Freedman, L.S., Ballard-Barbash, R., Troiano, R.P., Bingham, S., Schoeller, D.A., Schatzkin, A., Carroll, R.J.: Structure of dietary measurement error: results of the open biomarker study. American journal of epidemiology **158**(1), 14–21 (2003)
7. Gleser, L.: Improvements of the naive approach to estimation in nonlinear errors-in-variables regression models. Contemp. Math **112**, 99–114 (1990)
8. Day, N.E., McKeown, N., Wong, M.-Y., Welch, A., Bingham, S.: Epidemiological assessment of diet: a comparison of a 7-day diary with a food frequency questionnaire using urinary markers of nitrogen, potassium and sodium. International journal of epidemiology **30**(2), 309–317 (2001)
9. Subar, A.F., Kipnis, V., Troiano, R.P., Midthune, D., Schoeller, D.A., Bingham, S., Sharbaugh, C.O., Trabulsi, J., Runswick, S., Ballard-Barbash, R., *et al.*: Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: the open study. American journal of epidemiology **158**(1), 1–13 (2003)
10. Natarajan, L., Pu, M., Fan, J., Levine, R.A., Patterson, R.E., Thomson, C.A., Rock, C.L., Pierce, J.P.: Measurement error of dietary self-report in intervention trials. American journal of epidemiology **172**(7), 819–827 (2010)
11. Kipnis, V., Freedman, L.S., Carroll, R.J., Midthune, D.: A bivariate measurement error model for semicontinuous and continuous variables: Application to nutritional epidemiology. Biometrics **72**(1), 106–115 (2016)
12. Kaaks, R., Riboli, E., van Staveren, W.: Calibration of dietary intake measurements in prospective cohort studies. American Journal of Epidemiology **142**(5), 548–556 (1995)
13. Agogo, G.O., van der Voet, H., van't Veer, P., Ferrari, P., Muller, D.C., Sánchez-Cantalejo, E., Bamia, C., Braaten, T., Knüppel, S., Johansson, I., *et al.*: A method for sensitivity analysis to assess the effects of measurement error in multiple exposure variables using external validation data. BMC medical research methodology **16**(1), 139 (2016)
14. Dellaportas, P., Stephens, D.A.: Bayesian analysis of errors-in-variables regression models. Biometrics, 1085–1095 (1995)
15. 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. Journal of biopharmaceutical statistics **27**(5), 741–755 (2017)
16. Lin, X.: A bayesian semiparametric accelerated failure time model for arbitrarily censored data with covariates subject to measurement error. Communications in Statistics-Simulation and Computation **46**(1), 747–756 (2017)
17. 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 & Data Analysis **113**, 177–193 (2017)

18. Rosner, B., Willett, W., Spiegelman, D.: Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. Statistics in medicine **8**(9), 1051–1069 (1989)
19. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2018). R Foundation for Statistical Computing. https://www.R-project.org/
20. AC, V.H.: 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). http://dx.doi.org/doi:10.14749/1472711307
21. Barnabas, R.V., van Rooyen, H., Tumwesigye, E., Murnane, P.M., Baeten, J.M., Humphries, H., Turyamureeba, B., Joseph, P., Krows, M., Hughes, J.P., *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. The lancet HIV **1**(2), 68–76 (2014)
22. Freedman, L.S., Schatzkin, A., Midthune, D., Kipnis, V.: Dealing with dietary measurement error in nutritional cohort studies. Journal of the National Cancer Institute **103**(14), 1086–1092 (2011)
23. Hoff, P.D.: A First Course in Bayesian Statistical Methods vol. 580. Springer, ??? (2009)
24. Plummer, M.: Rjags: Bayesian Graphical Models Using MCMC. (2018). R package version 4-8
25. Lunn, D., Spiegelhalter, D., Thomas, A., Best, N.: The bugs project: Evolution, critique and future directions. Statistics in medicine **28**(25), 3049–3067 (2009)
26. Martin, A.D., Quinn, K.M., Park, J.H.: MCMCpack: Markov chain monte carlo in R. Journal of Statistical Software **42**(9), 22 (2011). http://www.jstatsoft.org/v42/i09/
27. Plummer, M., Best, N., Cowles, K., Vines, K.: Coda: convergence diagnosis and output analysis for mcmc. R news **6**(1), 7–11 (2006)
28. Plummer, M., Stukalov, A., Denwood, M., Plummer, M.M.: Package 'rjags'. update **16**, 1 (2018)
29. Feskanich, D., Rimm, E.B., Giovannucci, E.L., Colditz, G.A., Stampfer, M.J., Litin, L.B., Willett, W.C.: Reproducibility and validity of food intake measurements from a semiquantitative food frequency questionnaire. Journal of the American Dietetic Association **93**(7), 790–796 (1993)
30. 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. Journal of epidemiology and biostatistics **4**(4), 285–295 (1999)
31. Eliopoulos, C., Klein, J., Koren, G.: Validation of self-reported smoking by analysis of hair for nicotine and cotinine. Therapeutic drug monitoring **18**(5), 532–536 (1996)
32. Secker-Walker, R.H., Vacek, P.M., Flynn, B.S., Mead, P.B.: Exhaled carbon monoxide and urinary cotinine as measures of smoking in pregnancy. Addictive behaviors **22**(5), 671–684 (1997)
33. Stram, D.O., Huberman, M., Wu, A.H.: Is residual confounding a reasonable explanation for the apparent protective effects of beta-carotene found in epidemiologic studies of lung cancer in smokers? American journal of epidemiology **155**(7), 622–628 (2002)
34. Fisher, R.A.: Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika **10**(4), 507–521 (1915)
35. Fisher, R.A.: On the'probable error'of a coefficient of correlation deduced from a small sample. Metron **1**, 1–32 (1921)
36. Ham, E., Kim, H.-J.: Evaluation of fruit intake and its relation to body mass index of adolescents. Clinical nutrition research **3**(2), 126–133 (2014)
37. Sharma, S.P., Chung, H.J., Kim, H.J., Hong, S.T.: Paradoxical effects of fruit on obesity. Nutrients **8**(10), 633 (2016)
38. Heid, I., Küchenhoff, H., Miles, J., Kreienbrock, L., Wichmann, H.: Two dimensions of measurement error: classical and berkson error in residential radon exposure assessment. Journal of Exposure Science and Environmental Epidemiology **14**(5), 365 (2004)

Figures

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

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

Tables

Table 1 Comparison of posterior Mean (Standard Deviation), posterior Median and 95% CI 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

	Estimate for fruit intake			**Estimate for vegetable intake**		
Method	Mean (SD)	Median	95% CI	Mean (SD)	Median	95% CI
Naive	0.009 (0.012)	0.009	(-0.014, 0.032)	0.008 (0.014)	0.008	(-0.018, 0.034)
Univariate	0.026 (0.036)	0.027	(-0.043,0.097)	0.031 (0.051)	0.031	(-0.070, 0.130)
Trivariate (Proposed)	0.026 (0.036)	0.026	(-0.044, 0.097)	0.033 (0.051)	0.033	(-0.069, 0.132)
	Estimate for smoking					
Model	Mean (SD)	Median	95% CI			
Naive	-0.253 (0.640)	-0.247	(-1.484, 0.971)			
Univariate	-0.740 (1.874)	-0.721	(-4.342, 2.841)			
Trivariate (Proposed)	-0.714 (1.875)	-0.695	(-4.323, 2.864)			

Table 2 The Mean (Standard Deviation), Median and 95% CI 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 (proposed) method in the sensitivity analysis by equating the limits of literature reported validity coefficients to different CIs

	Estimate for fruit intake			**Estimate for vegetable intake**		
CI (%)	Mean (SD)	Median	95% CI	Mean (SD)	Median	95% CI
85	0.027 (0.038)	0.027	(-0.046, 0.101)	0.032 (0.051)	0.032	(-0.068, 0.131)
90	0.026 (0.037)	0.027	(-0.045,0.099)	0.032 (0.051)	0.032	(-0.068, 0.132)
95	0.026 (0.036)	0.026	(-0.044, 0.097)	0.033 (0.051)	0.033	(-0.069, 0.132)
99	0.025 (0.035)	0.025	(-0.043, 0.094)	0.033 (0.052)	0.033	(-0.069, 0.133)
	Estimate for smoking					
CI (%)	Mean (SD)	Median	95% CI			
85	-0.695 (1.839)	-0.676	(-4.237, 2.816)			
90	-0.704 (1.856)	-0.685	(-4.278, 2.838)			
95	-0.714 (1.875)	-0.695	(-4.323, 2.864)			
99	-0.727 (1.899)	-0.708	(-4.386, 2.897)			

Table 3 The Mean (Standard Deviation), Median and 95% CI 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 (proposed) method in the sensitivity analysis by varying the magnitude of error correlation between measurements for fruit and vegetable

	Estimate for fruit intake			**Estimate for vegetable intake**		
$\rho_{\epsilon_{W_1}\epsilon_{W_2}}$	Mean (SD)	Median	95% CI	Mean (SD)	Median	95% CI
0.00	0.019 (0.062)	0.019	(-0.103, 0.140)	0.024 (0.084)	0.023	(-0.139, 0.189)
0.05	0.021 (0.049)	0.021	(-0.076, 0.115)	0.026 (0.067)	0.025	(-0.105, 0.158)
0.10	0.022 (0.042)	0.023	(-0.060,0.103)	0.028 (0.058)	0.028	(-0.086, 0.142)
0.15	0.024 (0.038)	0.025	(-0.049, 0.098)	0.031 (0.053)	0.031	(-0.075, 0.133)
0.20	0.027 (0.036)	0.027	(-0.042, 0.097)	0.034 (0.051)	0.034	(-0.065, 0.132)
0.25	0.030 (0.036)	0.030	(-0.039, 0.100)	0.038 (0.051)	0.038	(-0.061, 0.137)
	Estimate for smoking					
$\rho_{\epsilon_{W_1}\epsilon_{W_2}}$	Mean (SD)	Median	95% CI			
0.00	-0.790 (1.868)	-0.776	(-4.413, 2.784)			
0.05	-0.773 (1.868)	-0.755	(-4.384, 2.801)			
0.10	-0.754 (1.870)	-0.738	(-4.361, 2.824)			
0.15	-0.731 (1.872)	-0.716	(-4.339, 2.840)			
0.20	-0.702 (1.876)	-0.683	(-4.319, 2.882)			
0.25	-0.667 (1.881)	-0.648	(-4.297, 2.936)			

Table 4 The Mean (Standard Deviation), Median and 95% CI 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 (proposed) method in the sensitivity analysis by varying the magnitude of error correlation between measurements for fruit and average number of cigarettes smoked

	Estimate for fruit intake			**Estimate for vegetable intake**		
$\rho_{\epsilon_{W_1}\epsilon_{W_3}}$	Mean (SD)	Median	95% CI	Mean (SD)	Median	95% CI
-0.25	0.033 (0.040)	0.034	(-0.045, 0.113)	0.029 (0.052)	0.029	(-0.074, 0.130)
-0.20	0.030 (0.038)	0.030	(-0.043, 0.105)	0.031 (0.052)	0.031	(-0.072, 0.130)
-0.15	0.028 (0.036)	0.028	(-0.043,0.099)	0.032 (0.052)	0.032	(-0.070, 0.131)
-0.10	0.026 (0.036)	0.026	(-0.044, 0.097)	0.032 (0.052)	0.033	(-0.069, 0.132)
-0.05	0.025 (0.037)	0.025	(-0.046, 0.098)	0.033 (0.051)	0.033	(-0.068, 0.132)
0.00	0.024 (0.039)	0.025	(-0.050, 0.101)	0.034 (0.051)	0.034	(-0.067, 0.133)
	Estimate for smoking					
$\rho_{\epsilon_{W_1}\epsilon_{W_3}}$	Mean (SD)	Median	95% CI			
-0.25	-1.259 (2.107)	-1.247	(-5.381, 2.833)			
-0.20	-1.060 (1.964)	-1.052	(-4.890, 2.74)			
-0.15	-0.899 (1.889)	-0.887	(-4.588, 2.711)			
-0.10	-0.744 (1.870)	-0.726	(-4.356, 2.845)			
-0.05	-0.620 (1.910)	-0.601	(-4.334, 3.037)			
0.00	-0.497 (2.011)	-0.487	(-4.402, 3.374)			

Table 5 The Mean (Standard Deviation), Median and 95% CI 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 (proposed) method in the sensitivity analysis by varying the magnitude of error correlation between measurements for vegetable and average number of cigarettes smoked

	Estimate for fruit intake			**Estimate for vegetable intake**		
$\rho_{\epsilon_{W_2}\epsilon_{W_3}}$	Mean (SD)	Median	95% CI	Mean (SD)	Median	95% CI
-0.28	0.023 (0.037)	0.024	(-0.049, 0.095)	0.044 (0.059)	0.044	(-0.071, 0.160)
-0.23	0.024 (0.037)	0.025	(-0.047, 0.096)	0.039 (0.054)	0.039	(-0.066, 0.146)
-0.18	0.025 (0.036)	0.026	(-0.045,0.096)	0.035 (0.051)	0.035	(-0.065, 0.135)
-0.13	0.026 (0.036)	0.026	(-0.044, 0.097)	0.033 (0.051)	0.033	(-0.067, 0.132)
-0.08	0.026 (0.036)	0.026	(-0.044, 0.097)	0.032 (0.053)	0.032	(-0.072, 0.134)
0.00	0.027 (0.036)	0.027	(-0.043, 0.097)	0.031 (0.061)	0.031	(-0.089, 0.150)
	Estimate for smoking					
$\rho_{\epsilon_{W_2}\epsilon_{W_3}}$	Mean (SD)	Median	95% CI			
-0.28	-1.361 (2.183)	-1.326	(-5.649, 2.846)			
-0.23	-1.120 (1.984)	-1.099	(-5.028, 2.723)			
-0.18	-0.912 (1.876)	-0.907	(-4.573, 2.689)			
-0.13	-0.744 (1.752)	-0.734	(-4.328, 2.804)			
-0.08	-0.603 (1.939)	-0.587	(-4.365, 3.105)			
0.00	-0.362 (2.295)	-0.354	(-4.779, 4.139)			

Figures

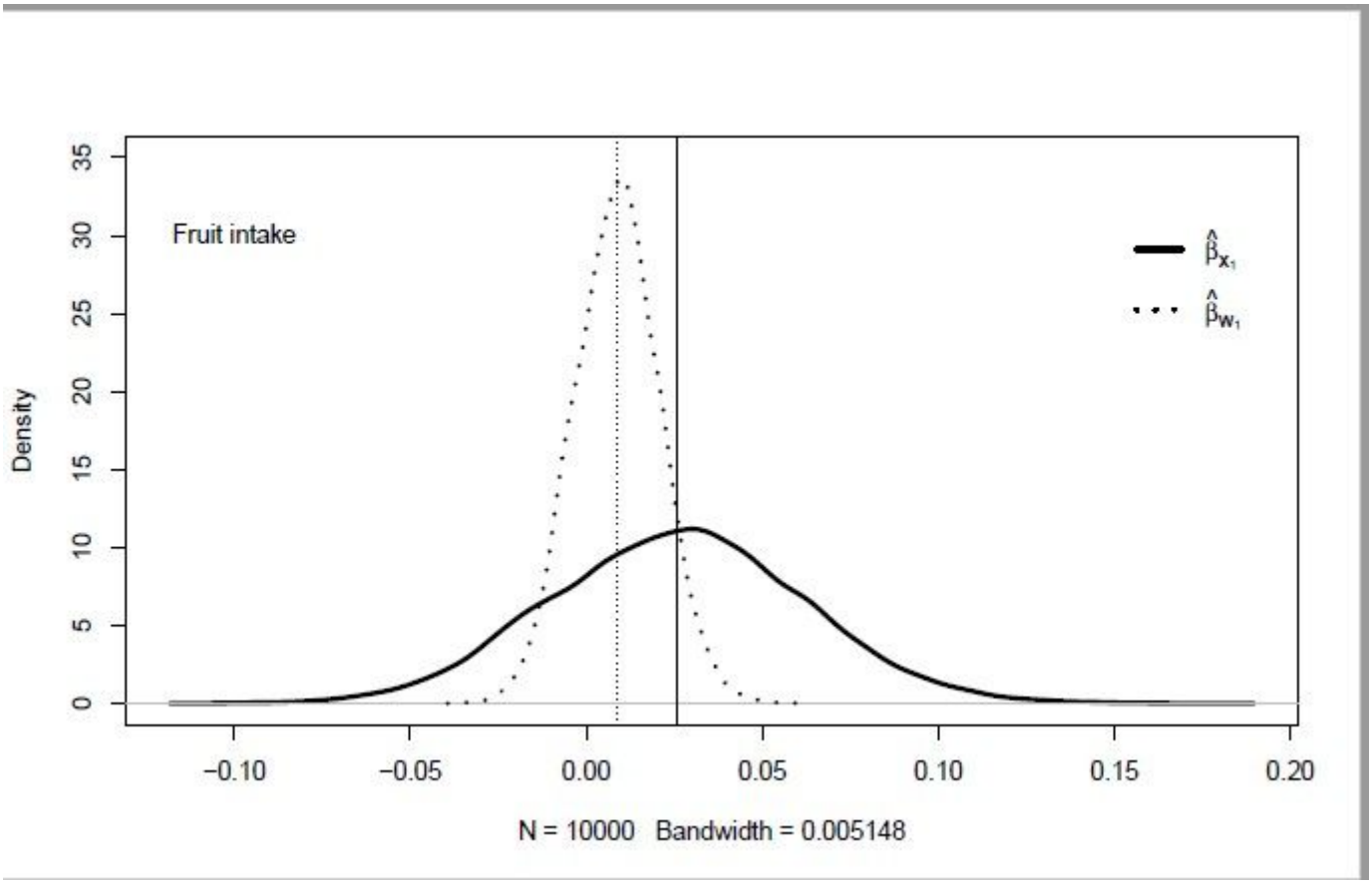

Figure 1

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 coeffcient estimates adjusted for bias; the dotted vertical lines indicate the posterior means of unadjusted coeffcient estimates

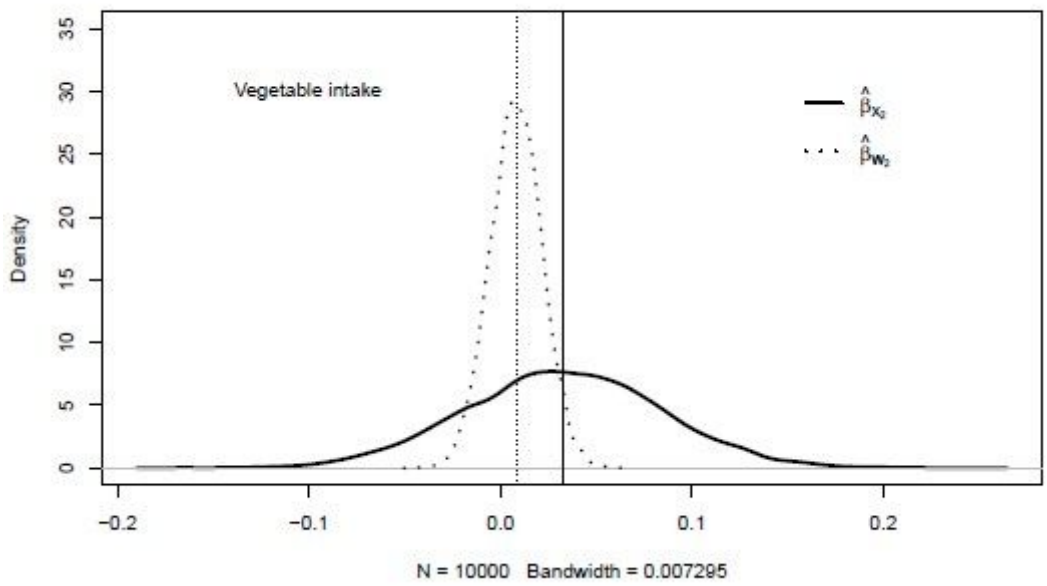

Figure 2

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

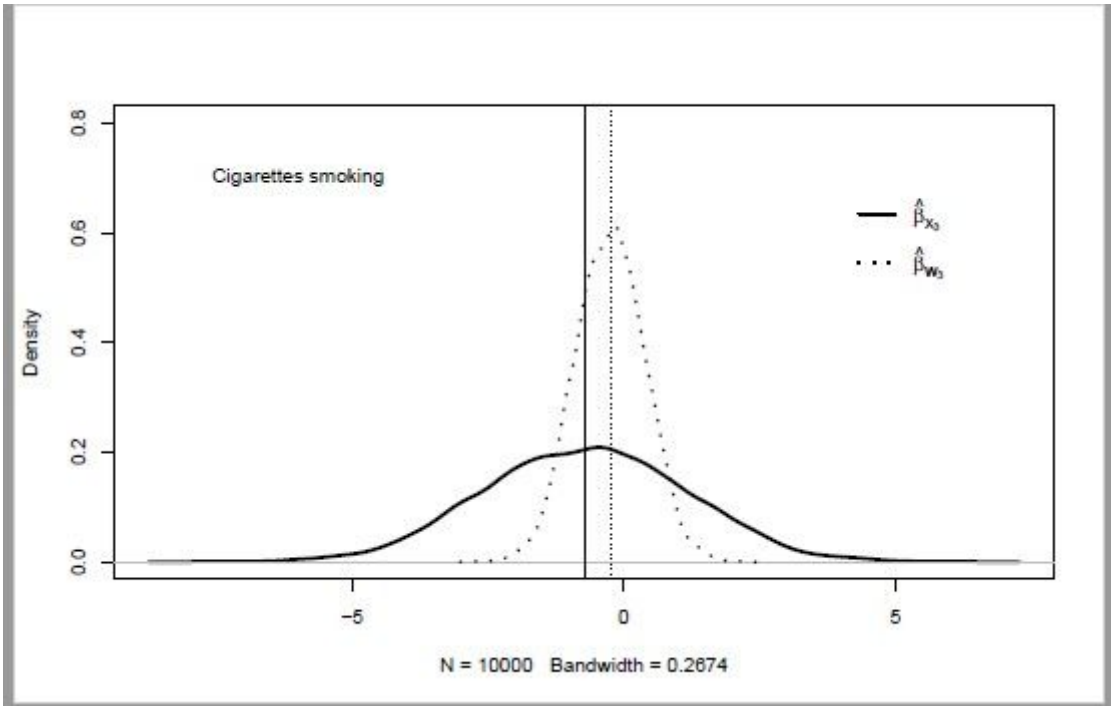

Figure 3

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

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.pdf
- referenceT.bib