Statistical Modelling for Exposure Measurement Error with Application to Epidemiological Data

George O. Agogo

Thesis Committee

Promotors

Prof. dr. Hendriek C. Boshuizen
Professor in Biostatistic Modelling in Nutrition Research
Division of Human Nutrition, Wageningen University

Prof. dr. Fred A. van Eeuwijk
Professor of Applied Statistics
Biometris, Wageningen University

Co-promotor

Dr. Hilko van der Voet
DLO Researcher, Biometris
Wageningen University and Research Centre

Other members

Prof. dr. ir. E.J.M. Feskens, Wageningen University
Dr. S. Le Cessie, Leiden University
Prof. dr. J. Wallinga, Leiden University
Dr. R.H. Keogh, London School of Hygiene and Tropical Medicine

This research was conducted under the auspices of the Graduate School VLAG (Advanced studies in Food Technology, Agrobiotechnology, Nutrition and Health Sciences).

Statistical Modelling for Exposure Measurement Error with Application to Epidemiological Data

George O. Agogo

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Thursday 14 January 2016
at 11:00 a.m. in the Aula.

George O. Agogo
Statistical Modelling for Exposure Measurement Error with Application to Epidemiological Data, 160 pages.

PhD thesis, Wageningen University, Wageningen, NL (2016)
With references, with summaries in English
ISBN 978-94-6257-622-3

Abstract

Background Measurement error in exposure variables is an important issue in epidemiological studies that relate exposures to health outcomes. Such studies, however, usually pay limited attention to the quantitative effects of exposure measurement error on estimated exposure-outcome associations. Therefore, the estimators for exposure-outcome associations are prone to bias. Existing methods to adjust for the bias in the associations require a validation study with multiple replicates of a reference measurement. Validation studies with multiple replicates are quite costly and therefore, in some cases only a single–replicate validation study is conducted besides the main study. For a study that does not include an internal validation study, the challenge in dealing with exposure measurement error is even bigger. The challenge is how to use external data from other similar validation studies to adjust for the bias in the exposure-outcome association. In accelerometry research, various accelerometer models have currently been developed. However, some of these new accelerometer models have not been properly validated in field situations. Despite the widely recognized measurement error in the accelerometer, some accelerometers have been used to validate other instruments, such as physical activity questionnaires, in measuring physical activity. Consequently, if an instrument is validated against the accelerometer, and the accelerometer itself has considerable measurement error, the observed validity in the instrument being validated will misrepresent the true validity.

Methodology In this thesis, we adapted regression calibration to adjust for exposure measurement error for a single-replicate validation study with zero-inflated reference measurements and assessed the adequacy of the adapted method in a simulation study. For the case where there is no internal validation study, we showed how to combine external data on validity for self-report instruments with the observed questionnaire data to adjust for the bias in the associations caused by measurement error in correlated exposures. In the last part, we applied a measurement error model to assess the measurement error in physical activity as measured by an accelerometer in free-living individuals in a recently concluded validation study.

Results The performance of the proposed two-part model was sensitive to the form of continuous independent variables and was minimally influenced by the correlation between the probability of a non-zero response and the actual non-zero response values. Reducing the number of covariates in the model seemed beneficial, but was not critical in large-sample studies. We showed that if the confounder is strongly linked with the outcome, measurement error in the confounder can be more influential than measurement error in the exposure in causing the bias in the exposure-outcome association, and that the bias can be in any direction. We further showed that when accelerometers are used to monitor the level of physical activity in free-living individuals, the mean level of physical activity would be underestimated, the associations between physical activity and health outcomes would be biased, and there would be loss of statistical power to detect associations.

Conclusion The following remarks were made from the work in this thesis. First, when only a single-replicate validation study with zero-inflated reference measurements is available, a correctly specified regression calibration can be used to adjust for the bias in the exposure-outcome associations. The performance of the proposed calibration model is influenced more by the assumption made on the form of the continuous covariates than the form of the response distribution. Second, in the absence of an internal validation study, carefully extracted validation data that is transportable to the main study can be used to adjust for the bias in the associations. The proposed method is also useful in conducting sensitivity analyses on the effect of measurement errors. Lastly, when “reference” instruments are themselves marred by substantial bias, the effect of measurement error in an instrument being validated can be seriously underestimated.

Table of Contents

1 Introduction

1.1 Measurement error in exposure variables

Measurement error in exposure variables is a known problem in many research areas. By definition, measurement error refers to the discrepancy between the true value and measured value of a variable (Thomas, Stram and Dwyer, 1993). In epidemiological research, measurement error can be due to recall bias when studies are conducted retrospectively requiring an individual to recall and report past events, or due to biological variations and instrument errors in laboratory experiments. Measurement error in exposures has been a long-term concern in relating exposures to health outcomes in epidemiological studies (Rothman, Greenland and Lash, 2008). In nutritional epidemiology, measurement error in dietary exposures has been a major impediment in relating long-term dietary intake to the occurrence of a disease (Ferrari et al., 2008). For example, the interest might be in the association between long-term intake of fruit and vegetable (hereafter, FV) and overall risk of cancer, as is the case in Boffetta et al. (2010). In such studies, dietary intake is usually assessed with a self-report instrument, which is prone to measurement error. Measurement error in the self-report instrument can be due to memory failure to recall past intakes over a long period of time (Agudo, 2004; Rosner and Gore, 2001). Therefore, it is impossible to measure long-term intake exactly.

Using a dietary example, measurement error in dietary exposures can have two important effects in the parameter estimate that relates dietary intake to occurrence of a disease (Carroll et al., 2006). First, measurement error in dietary exposure can bias the parameter estimate that quantifies the association between dietary intake and occurrence of a disease (hereafter, diet-disease association). Second, there can be loss of statistical power to detect existing diet-disease associations. To illustrate these two important effects of measurement error, we use simulated data on the association between FV intake and reduction in the risk of cancer as an example (Figure 1.1).

Random measurement error in FV intake leads to scatterplots (diamond dots) that are either higher or lower than true values (round dots). As a result, the regression slope that quantifies the association between FV intake and lower risk of cancer becomes more flattened (solid line) than the true slope (dashed line). This flattening effect quantifies the attenuation due to measurement error in the exposure. Attenuation refers to the bias of the association toward the null (Kipnis et al., 2003). Additionally, the variability around the regression line is much greater when the measured intake is used (diamond dots) than the variability around the regression line when the true intake is used (round dots), demonstrating *loss of statistical power*. The implication of the loss of power is that larger sample sizes are required to detect associations.

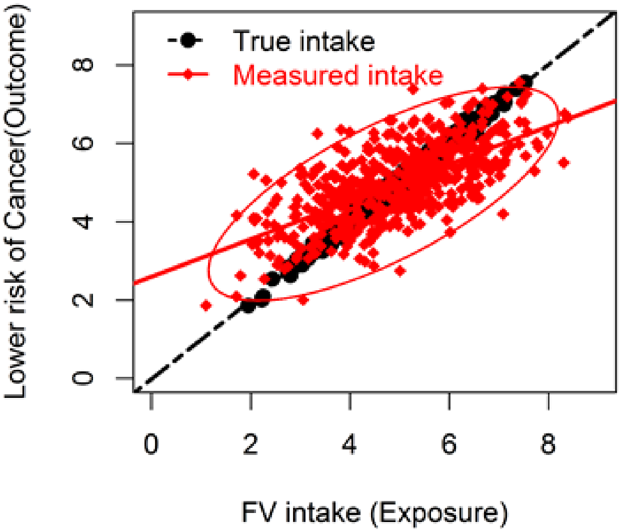

Figure 1.1: Simulated data showing the two important effects of measurement error on the association between reduction of risk of cancer (outcome) and FV intake (exposure): bias in the exposure-outcome association (attenuation) and loss of statistical power to detect associations. True intake (exposure) is shown by round dots and measured intake by diamond dots.

The focus of this thesis is on methods to adjust for the bias in the exposure-outcome associations, when exposure variables are measured with errors. When multiple exposures are measured with correlated errors, the exposure-outcome association can be biased in any direction (Day et al., 2004; Michels et al., 2004; Rosner et al., 2008). The exposure measurement error problem has prompted much methodological research, initially, on understanding the effects of measurement error on exposure-outcome associations and, more recently, on developing statistical methods to correct for exposure measurement error (Buonaccorsi, 2010; Carroll et al., 2006).

1.2 Measurement error in dietary exposure assessment

The commonly used self-report instruments for measuring dietary intakes include food dietary questionnaires (DQs), 24-hour dietary recalls (24HRs) and dietary records (Agudo, 2004; Black, Welch and Bingham, 2000). In measuring dietary intake with a DQ, a wide range of food items are listed and individuals are asked to report their average frequency of consumption and average amount consumed over a long period of time, typically between several months to a year. A major drawback associated with the use of the DQ is memory failure to recall past intake. In the 24HR, however, subjects are asked to report their intake within 24 hours prior to the assessment time. Similar to the DQ, 24HR is also prone to recall bias, but to a lesser extent. Additionally, a single 24HR is unreliable in measuring long-term intake, unless administered multiple times per individual (Agudo, 2004). Unlike DQ and 24HR, in dietary records, an individual keeps a record of actual intake of foods consumed at the time of consumption for a specified period (Willet, 1998). Measuring food intake with the dietary record can lead to individuals changing their food intake patterns, when they realize that their intake is being monitored. This could lead to misreporting of true intake in the dietary record (Johnson, 2002). Further details on dietary assessment methods can be found in Al-Delaimy et al. (2005), Bingham et al. (1997), Bingham et al. (1994) and Block (1982, 1989).

1.3 Common measurement error terminologies, types and structures

Measurement error can either be *systematic* or *random*. Systematic error arises when an individual consistently overestimates or underestimates his true level of exposure

(Agudo, 2004). Random error, however, arises when an individual sometimes overestimates and sometimes underestimates true level of exposure.

Exposure measurement error can either be *nondifferential* or *differential*. Nondifferential error occurs when the measured exposure contains no extra information about the health outcome other than what is contained in true exposure; error is said to be differential otherwise (Carroll et al., 2006, p. 36). In the FV intake-cancer example, if measurement error is nondifferential, measured FV intake in a single day should not contribute extra information about the risk of cancer other than what is contained in true long-term FV intake. On the other hand, if measurement error in reported FV intake depends upon whether an individual has cancer or not, especially due to recall bias when such a study is conducted retrospectively, then the measurement error will be differential (Thomas et al., 1993).

To define different types of measurement error using the FV intake example, we denote self-report intake by Q, unknown long-term true intake by T and measurement error in self-reported intake by ε_Q.

Measurement error structure can either be of *classical* or *Berkson* type. Classical error occurs when true intake is measured with error, such that the variability in the measured intake is greater than the variability in true intake. An example of classical measurement error structure is given by

$$Q = T + \varepsilon_Q. \tag{1.1}$$

Berkson error, on the other hand, occurs when true intake is the sum of measured intake and measurement error, such that the variability in true intake is greater than the variability in the measured intake (Berkson, 1950). An example of Berkson error structure is given by

$$T = Q + \varepsilon_{Q.} \tag{1.2}$$

An example where Berkson error occurs is when individuals are classified into groups and all group members are assigned the same exposure value, say, the mean of true intake within that group; however, true intake for individuals within the same group would differ from the assigned intake values (Thomas et al., 1993). Exposures can also be measured with a mixture of Berkson and classical error as shown in Carroll et al. (2006, p. 51).

The structure of measurement error can either be *additive* (as in expression (1.1)) or *multiplicative*. Multiplicative error arises when the magnitude of error increases with the quantity of true intake, such that larger true intake values are measured with larger error (Guolo and Brazzale, 2008). Multiplicative error is usually common with dietary intakes (Carroll et al., 2006). An example of a multiplicative measurement error structure can be given as

$$Q = T\varepsilon_Q. \tag{1.3}$$

Exposure measurements can either be *unbiased* (e.g., as shown in expression (1.1)) or *biased*. Intake measurements are biased if the average value does not equal the true mean value, meaning that there is systematic error in Q.

With multiple replicate self-report measurements (denoted by *j* for replicate and *i* for an individual), an example of additive measurement error structure with both systematic and random error components can be given by

$$Q_{ij} = \beta_0 + \beta_Q T_i + r_{Q_i} + \varepsilon_{Q_{ij}}, \tag{1.4}$$

where β_0 quantifies overall constant bias and β_Q quantifies overall bias that is related to the level of intake (Kipnis et al., 2001); r_{Q_i} is referred to as person-specific bias and describes the fact that two individuals who consume the same amount of food will systematically report their intakes differently (Carroll et al., 2006). For further details on measurement error structures, see Buonaccorsi (2010), Carroll et al. (2006) and Fuller (2006).

In Figure 1.2, systematic and random measurement errors are juxtaposed; also shown are the components of an additive measurement error structure that is shown in expression (1.4).

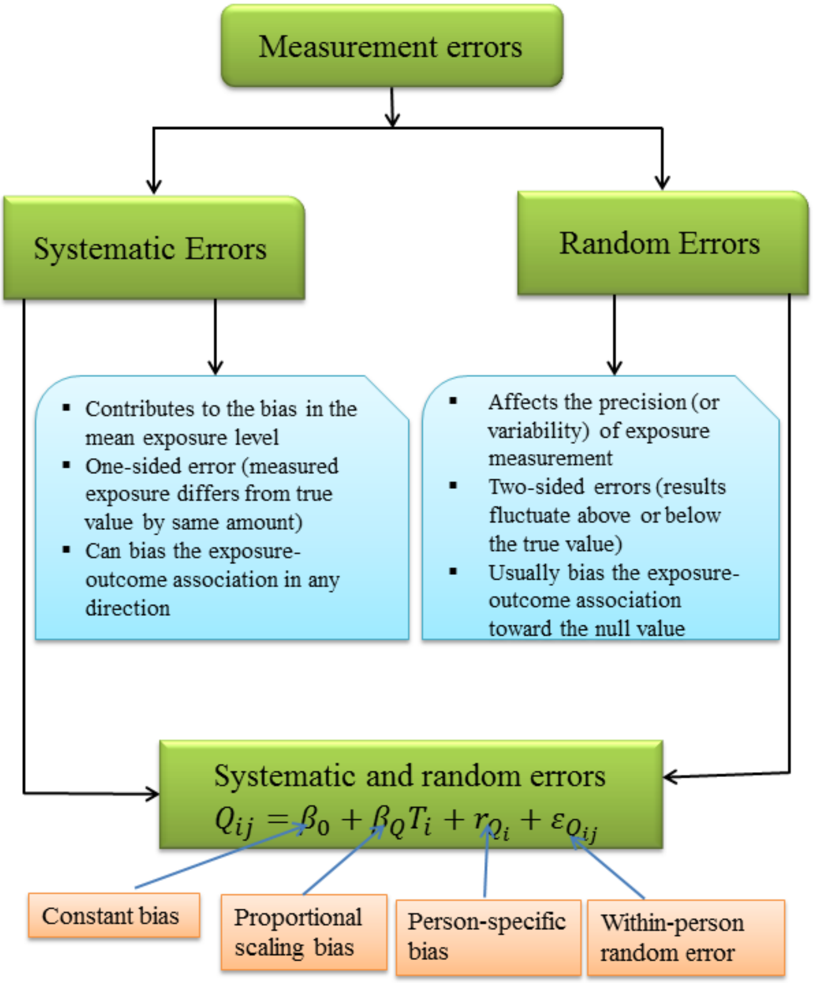
Measurement errors
Systematic Errors
Random Errors
Contributes to the bias in the mean exposure level
One-sided error (measured exposure differs from true value by same amount)
Can bias the exposure-outcome association in any direction
Affects the precision (or variability) of exposure measurement
Two-sided errors (results fluctuate above or below the true value)
Usually bias the exposure-outcome association toward the null value
Systematic and random errors
$Q_{ij} = \beta_0 + \beta_Q T_i + r_{Q_i} + \varepsilon_{Q_{ij}}$
Constant bias
Proportional scaling bias
Person-specific bias
Within-person random error

Figure 1.2 Schematic representation of common measurement error structures: systematic, random and both errors.

In Figure 1.3, the additive measurement error structures are presented graphically using an example of simulated intake data for an individual. In each graph, the solid horizontal lines represent an individual's usual intake, defined as the average intake for the ten consumption days. When daily intake is measured with random error only, usual intake is measured accurately, i.e., with no bias as shown in Figure 1.3(a), but imprecisely leading to an inflated variance as shown by wider width of the green solid density curve in Figure 1.3(d). Figure 1.3(b) shows a situation where intake is measured with systematic error (here, constant bias only), and in this case intake is consistently overestimated with a distribution that is shifted to the right of the true distribution, but with the same variance as for the true intake (Figure 1.3(d), blue solid curve). In other cases, however, systematic error can lead to consistent underestimation of intake. Further, when intake is measured with both systematic and random errors, daily intake might be sometimes overestimated and sometimes underestimated, leading to an upward bias in the usual intake as shown in Figure 1.3(c) and an inflated variance as shown in Figure 1.3(d), red solid curve. In other cases, however, where intake is measured with both systematic and random errors, usual intake can be underestimated or overestimated. The wavy lines represent within-individual variation in daily intakes (Figure 1.3(a), (b) and (c)).

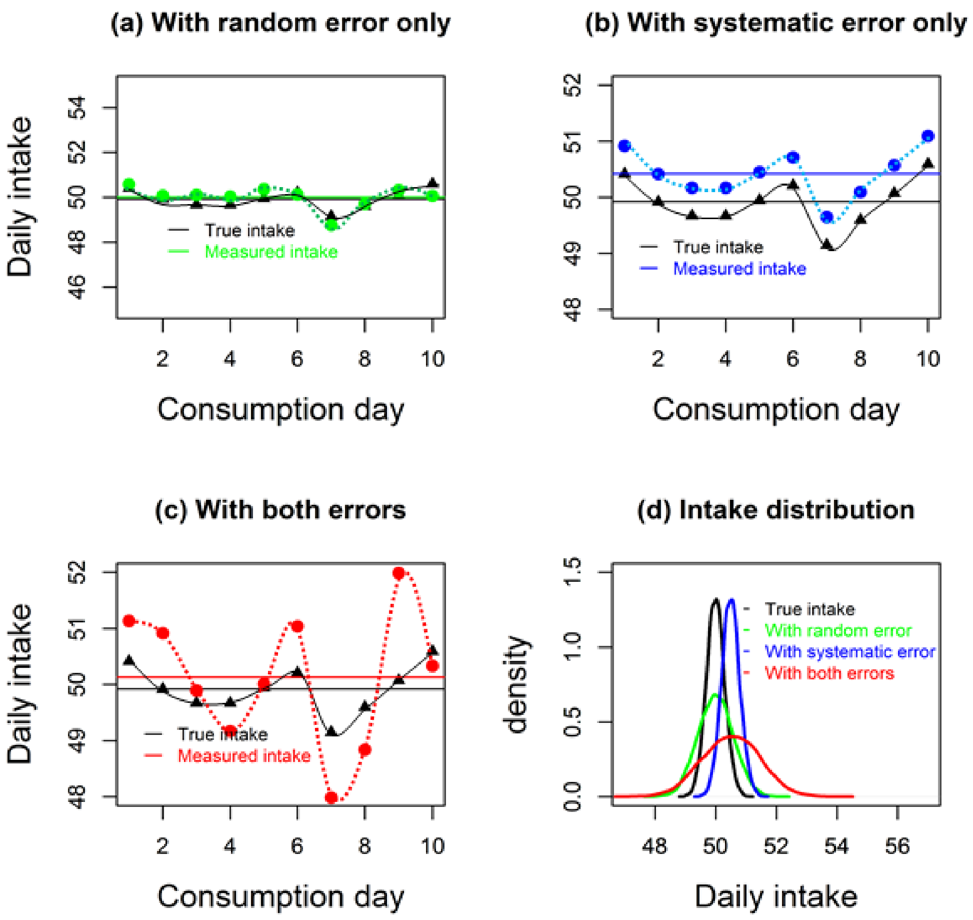

Figure 1.3: A graphical representation of three common types of additive measurement error structures using simulated data when daily intake is measured with random error only, systematic error only and both systematic and random error. The horizontal lines represent usual intake; the black dots represent true intake values per consumption day; the dots in other colours represent measured intake values. Figure 1.3(d) compares an individual's daily intake distributions for true intake and measured intakes.

1.4 Common methods to correct for measurement error in continuous exposures

Five common bias-adjustment methods are described in this section. They are summarized in Table 1.1.

1.4.1 Regression calibration

Regression calibration is the most commonly used method to adjust for the bias in the exposure-outcome associations caused by measurement errors in the exposure variables (Freedman et al., 2008; Guolo and Brazzale, 2008; Messer and Natarajan, 2008; Rosner, Spiegelman and Willett, 1990; Rosner, Willett and Spiegelman, 1989). This method is widely used because of its simplicity. The general idea of regression calibration is to estimate the conditional expectation of true exposure given the exposure measured with error and other covariates assumed to be measured without error. The estimated expected values are used in place of the unknown true exposure to estimate the association between the exposure and an outcome. Application of regression calibration requires additional information on the unobserved true exposure. This information is usually obtained from a validation study with unbiased measurements for true exposure (Kipnis et al., 1999; Kipnis et al., 2003). A validation study is often smaller than the main study and can be a random sample of subjects in the main study. To apply regression calibration, measurement error in the exposure is assumed to be nondifferential (Keogh and White, 2014). The method usually leads to consistent estimators of the association parameter (Guolo and Brazzale, 2008).

1.4.2 Likelihood method

The use of the likelihood method to correct for exposure measurement error requires specification of a parametric model for every component of the data, namely, the relation between the outcome of interest and the true exposure (hereafter, disease model), the relation between true exposure and covariates such as age and BMI (hereafter, exposure model) and the relation between measured exposure and the true exposure (hereafter, measurement error model). Subsequently, the likelihood function is obtained by integrating the product of the three densities over the latent true exposure variable (Guolo and Brazzale, 2008). The likelihood function is maximized to estimate the model parameters using either numerical methods or analytical approximations (Guolo and Brazzale, 2008). The method requires a validation study with multiple replicates of the unbiased measurements in order to specify the distribution of the true exposure. The likelihood method assumes nondifferential error

(Thoresen and Laake, 2000). If correctly specified, the likelihood method can be more efficient than simpler measurement error correction methods such as regression calibration. This method, however, is rarely used in practice due to its computational burden and difficulty to specify parametric assumptions correctly. Moreover, it is often difficult to understand robustness of the likelihood method to model assumptions (Carroll et al., 2006).

1.4.3 Simulation Extrapolation (SIMEX)

SIMEX is a simulation-based method, sharing the simplicity and generality of regression calibration, but can be computationally prohibitive (Carroll et al., 2006). The method was proposed by Cook and Stefanski (1994) and has been developed further (Carroll et al., 1996; Lin and Carroll, 1999; Stefanski and Cook, 1995). The SIMEX method requires knowledge of the measurement error variance. Thus, a validation study with at least two replicates of the unbiased measurements is required. The following steps are followed to adjust for the bias in the association using the SIMEX method (Carroll et al., 2006). First, extra measurement error is added to the exposure measurements, creating a dataset with larger measurement error variance; this data generation step is repeated a large number of times. Second, the disease model is fitted and the association parameter estimated from each of the generated datasets in step one, then the average of the association parameter estimate is computed. Third, the above steps are repeated by adding various magnitudes of extra measurement error. Fourth, the average parameter estimates are plotted against the magnitude of added measurement error variance and then an extrapolant function is fitted. Lastly, the extrapolation is done to the ideal case of no measurement error to obtain the bias-corrected estimate. The method assumes nondifferential error and is well suited for exposures measured with additive or multiplicative error (Guolo and Brazzale, 2008).

1.4.4 Bayesian methods

Bayesian methods have been used to adjust for measurement error in the covariates (Dellaportas and Stephens, 1995; Huang, Chen and Dagne, 2011). To adjust for the exposure measurement error with the Bayesian method, the following steps are followed (Carroll et al., 2006, p. 206-207). Similar to the likelihood method, a parametric model is specified for each component of the data and a likelihood function is formed. In the Bayesian framework, the parameters are assumed as random, unlike in the likelihood method where the parameters are assumed as fixed. Prior distributions are specified for the parameters in the model. Lastly, the posterior summary measures for the association parameter estimates are computed. The computation can be done with either a flexible sampling-based Markov chain Monte Carlo (MCMC) (Gelman and Hill, 2007) or a more computationally efficient non-sampling based integrated nested Laplace approximation (INLA) (Muff, 2015; Rue, Martino and Chopin, 2009). Despite the flexibility of Bayesian MCMC, it can be computationally intensive. Similar to the likelihood method, application of Bayesian method requires a validation study with multiple replicates of the unbiased measurements in order to specify the distribution of the true exposure. The method is suitable for nondifferential errors.

1.4.5 Multiple Imputation

Multiple imputation is a standard technique for handling data that are missing at random (Rubin, 1987). This method was proposed to adjust for measurement error in continuous exposures (Cole, Chu and Greenland, 2006; Freedman et al., 2008). To apply the multiple imputation method, the true exposure is assumed as missing data and is imputed multiple times by drawing from a distribution of the true exposure given all the observed data, including the outcome (Keogh and White, 2014). The multiple imputation method can accommodate differential error, because true exposure is imputed dependent on the outcome. The conditional distribution of the true exposure given the outcome and other observed data is often unknown. To estimate this conditional distribution, a validation study is required with data on (a) the study outcome and (b) multiple replicates of the unbiased measurements per individual. Using the estimated conditional distribution, the true exposure is imputed multiple times per individual to account for the uncertainty in the imputed exposure values. To each imputed dataset, the exposure-outcome model is fitted and the resulting association estimates are combined to obtain a pooled mean estimate (Rubin, 1987). The pooled mean estimate yields the bias-corrected estimator for the true exposure-outcome association (Messer and Natarajan, 2008).

Table 1.1: Summary details for common measurement error adjustment methods

Methods	Error assumption	Required data	Steps
Regression calibration	Nondifferential	• Validation study with unbiased exposure measurements	1) Conditional expectation of true exposure given observed data is estimated 2) The expected values are used in place of unknown true exposure to estimate the exposure-outcome association 3) Standard error of the association parameter can be estimated using either bootstrap or asymptotic methods
Likelihood	Nondifferential	• Validation study with at least two replicates of unbiased exposure measurements per subject	1) Parametric model is specified for each component of the data, i.e., the outcome model, exposure model and measurement error model 2) The likelihood function is formed and maximized
SIMEX	Nondifferential	• Validation study with at least two replicates of the unbiased exposure measurements per subject	1) Extra error is added to the measured exposure measurements to generate a new dataset 2) Association parameter is estimated from the generated dataset 3) The above steps are repeated many times and the mean estimate of the association parameter is computed 4) Steps1-3 are repeated for various magnitudes of extra-added error 5) The average association estimates are plotted against the magnitude of extra-added error and a trend is established 6) The trend is extrapolated back to the case of no error to obtain the SIMEX estimate
Bayesian	Nondifferential	• Validation study with at least two replicates of the unbiased measurements per subject • Prior distributions for the model parameters	1) Model is specified for each data component and the likelihood function is formed 2) Prior distribution is specified for each model parameter 3) Posterior distribution is estimated and summary measures computed for the association parameter estimate
Multiple imputation	Differential	• A validation study with (i) outcome data and (ii) at least two replicates of the unbiased exposure measurements per subject	1) True exposure is imputed multiple times by drawing from a distribution of the true exposure given all the observed data, including the outcome 2) The exposure-outcome model is fitted to each imputed dataset to estimate the association 3) The association estimates are combined to obtain pooled mean estimate

1.5 Current challenges

Most validation studies in nutritional research stop at describing the correlation of measured intake with true intake. Studies that look at the quantitative effects of measurement error in dietary intakes on estimated associations between intake and health outcomes are rare. Additionally, the effects of dietary intakes on health outcomes are usually weak and are marred by inconsistencies. These inconsistencies are partly due to measurement error in intake, because many nutritional studies are based on questionnaires or interviews that contain a large amount of measurement error.

Conducting a multiple-replicate validation study, besides the main study, however, is limited because it is costly. As a result, some epidemiological studies either conduct a single–replicate validation study or do not conduct a validation study at all. Among the commonly used measurement error correction methods (see Table 1.1) only regression calibration can be used for single-replicate calibration studies. However, regression calibration has not been applied and evaluated for a single-replicate validation study with zero-inflated measurements for the calibration response.

For a study that does not include an internal validation study, the challenge in dealing with exposure measurement error is even bigger. The challenge is how to use external validation data from other similar studies to adjust for the bias in the exposure-outcome association. When exposures are measured with correlated errors, it can be very difficult to predict the direction and strength of the association (Marshall, Hastrup and Ross, 1999). The difficulty is due to *contamination effect* of the confounder measurement error (Freedman et al., 2011). Even though the problem due to contamination effect has been widely acknowledged in the literature, there is a lack of practical methods to quantify this effect in a specific epidemiologic study, both in terms of the approximate magnitude of the effect, and its direction. Measurement error problem is also common in physical activity research, where instruments, such as physical activity questionnaires, physical activity recalls and accelerometers, are used to monitor an individual’s long-term level of physical activity (Ferrari, Friedenreich and Matthews, 2007; Hills, Mokhtar and Byrne, 2014; Lim et al., 2015; Nusser et al., 2012; Tooze et al., 2013). Regarding the accelerometry research, various accelerometer models have currently been developed. However, some of these new accelerometer models have not been properly validated in field situations. Despite the widely recognized measurement error in the accelerometer, some accelerometers have been used to validate other instruments, such as physical activity questionnaires, in measuring physical activity (Lim et al., 2015). Therefore, if an instrument is validated against the accelerometer, and the accelerometer itself has considerable measurement error, the observed validity in the instrument being validated will misrepresent the true validity.

These challenges constitute the motivation for the work in this thesis. The work in this thesis will address the following research questions emanating from the above-mentioned challenges:

(i) When only a single-replicate validation study with zero-inflated measurements is available, can the current methods be adapted to adjust for exposure measurement error?

(ii) When there is no internal validation study, can a practical method be proposed that uses external validation data to adjust for the bias in the exposure-outcome associations?

(iii) How large is the error in physical activity as measured by an accelerometer in free-living individuals and what is the impact of this error when accelerometer is used to validate other instruments?

1.6 Objectives of the study

The research in this thesis aims to address the aforementioned research questions. The key objectives of this thesis are highlighted below:

1) To propose a two-part regression calibration to adjust for measurement error in dietary intakes not consumed daily, when only a single-replicate validation study is

available. The task is to start with a simple linear calibration model and then improve it gradually. The improvement is done by modelling the excess zeros explicitly, handling heteroscedasticity in the response, exploring the optimal variable selection criteria, and identifying the optimal parametric forms of the continuous covariates in the calibration model,

2) To assess the performance of the proposed two-part regression calibration model in a simulation study with respect to: the percentage of excess zeroes in the response variable, the magnitude of correlation between probability of a non-zero response and the actual non-zero value, percentage of zeroes in the response and the magnitude of measurement error in the exposure,
3) To develop a multivariate method to adjust for the bias in the exposure-outcome association in the presence of mismeasured confounders when there is no internal validation study. The method combines external data on the validity of self-report instruments with the observed data to adjust for the bias in the exposure-outcome association, while simultaneously adjusting for confounding and measurement error in the confounders,
4) To validate a triaxial accelerometer in a recently concluded study by applying a measurement error model and quantifying the effects of measurement error in physical activity as measured by the accelerometer.

1.7 Thesis outline

This thesis is organized into chapters. The contents of the remaining chapters are summarized below.

In **chapter 2**, a two-part regression calibration model, initially developed for a multiple-replicate validation study design, is adapted to a case of a single-replicate validation study. The chapter further describes how to: handle the excess zeroes in the response, using two-part modelling approach; explore optimal parametric forms of the continuous covariates, using generalized additive modelling and empirical logit approaches, and how to select covariates into the calibration model. The adapted two-part model is compared with simple calibration models for episodically consumed food measured with error. A real epidemiologic case-study data is used.

In **chapter 3**, a simulation study is conducted to assess the performance of the proposed two-part regression calibration by mimicking the case-study data in chapter 2.

In **chapter 4**, a multivariate method is proposed to adjust for exposure measurement error, confounding and measurement error in the confounders when there is no internal validation study. The proposed method uses Bayesian Markov chain Monte Carlo method to combine prior information on the validity of self-reports with the observed data to adjust for the bias in the association. The method is compared with a method that ignores measurement error in the confounders. Further, a sensitivity analysis is performed to get insight into the measurement error structure, especially with respect to the magnitude of error correlation. The proposed method is illustrated with a real dataset.

In **chapter 5**, a triaxial accelerometer is validated against doubly labelled water using a proposed measurement error model. Measurement error in the accelerometer is quantified with: (a) the bias in the mean level of physical activity, (b) the correlation coefficient between measured and true physical activity to quantify loss of statistical power in detecting associations, and (c) attenuation factor to quantify the bias in the associations between physical activity and health outcomes.

In **chapter 6**, the main findings from the thesis are summarized and discussed in a general context. The study limitations are highlighted followed by suggestions for improvement and potential areas for future research. The chapter ends with concluding remarks.

2 Use of two-part regression calibration model to correct for measurement error in episodically consumed foods in a single-replicate study design: EPIC Case Study [1]

Abstract

In epidemiologic studies, measurement error in dietary variables often attenuates association between dietary intake and disease occurrence. To adjust for the attenuation, regression calibration is commonly used. To apply regression calibration, unbiased (reference) measurements are required. Short-term reference measurements for foods not consumed daily contain excess zeroes that pose challenges in the calibration model. We adapted two-part regression calibration model, initially developed for multiple replicates of reference measurements per individual to a single-replicate setting. We showed how to handle excess zero reference measurements by two-step modelling approach, how to explore heteroscedasticity in the consumed amount with variance-mean graph, how to explore nonlinearity with the generalized additive modelling (GAM) and the empirical logit approaches, and how to select covariates in the calibration model. The performance of two-part calibration model was compared with the one-part counterpart. We used vegetable intake and mortality data from European Prospective Investigation on Cancer and Nutrition (EPIC) study. In the EPIC, reference measurements were taken with 24-hour recalls. For each of the three vegetable subgroups assessed separately, correcting for error with an appropriately specified two-part calibration model resulted in about three fold increase in the strength of association with all-cause mortality, as measured by the log hazard ratio. Further found is that the standard way of including covariates in the calibration model can lead to over fitting. Moreover, the extent of adjusting for measurement error is influenced by forms of covariates in the calibration model.

[1] Based on: Agogo, G. O., van der Voet, H., van 't Veer, P., et al. (2014). Use of Two-Part Regression Calibration Model to Correct for Measurement Error in Episodically Consumed Foods in a Single-Replicate Study Design: EPIC Case Study. PLoS ONE 9, e113160.

2.1 Introduction

Dietary variables are often measured with error in nutritional epidemiology. In such studies, long-term dietary intake (usual) dietary intake is assessed with instruments such as food frequency questionnaire and dietary questionnaire (Agudo, 2004; Kaaks et al., 2002; Willet, 1998). In these instruments, the queried period of intake ranges from several months to a year, resulting in difficulties to recall past intake of foods or food groups, the frequency of consumption, and the portion size. In general, the measurement error in dietary intake can either be systematic or random. Systematic error occurs when an individual systematically overestimates or underestimates dietary intake, whereas random error is due to random within-individual variation in reporting of dietary intake (Kaaks et al., 2002; Kipnis et al., 2003). The random error attenuates the association between dietary intake and disease occurrence, whereas systematic error can either attenuate or inflate the association.

As a case study, we used the European Prospective Investigation on Cancer and Nutrition (EPIC) study. In EPIC, country-specific dietary questionnaires, hereafter DQ, were used to measure usual intake of various dietary variables or groups of dietary variables in different participating cohorts. With DQ measurements for usual intake, an association parameter estimate that relates usual intake to disease occurrence is often biased, typically towards the null (Fraser and Stram, 2001; Kaaks, 1997; Kipnis et al., 2003).

Regression calibration is the commonly used method to adjust for the bias in the association between usual intake and disease occurrence, due to measurement error in the DQ. Regression calibration involves finding the best conditional expectation of true intake given DQ intake and other error-free variables (Freedman et al., 2008). The mean expected intake values are used in place of true usual intake in a disease model that relates dietary intake to disease occurrence. Regression calibration requires a calibration sub-study to obtain unbiased intake measurements to be used as the calibration response. Some prospective studies therefore include a calibration sub-study that can either be internal or external. Internal calibration study consists of a random sample of subjects from the main study, as was the case in the EPIC, whereas external calibration sub-study consists of subjects not in the main study but with similar characteristics as the main-study subjects (Slimani et al., 2002). In the calibration sub-study, unbiased (or reference) measurements are collected by short-term instruments such as food records or 24-hour dietary recalls. In the EPIC study, regression calibration can also adjust for systematic error in the DQ due to the multicentre effect (Ferrari et al., 2008; Ferrari et al., 2004). In the EPIC calibration sub-study, a 24-hour dietary recall (hereafter, 24-HDR) was used as the reference instrument. From each subject in the EPIC calibration sub-study, only a single measurement was obtained (Slimani et al., 2002). Dietary intake reported in the 24HDR for food not consumed daily is usually characterized by excess zeroes. These excess zeroes pose a challenge in the calibration model (Kipnis et al., 2009; Olsen and Schafer, 2001; Tooze et al., 2006; Zhang et al., 2011). With regression calibration, the excess zeroes can be handled using a two-step approach, where in the first step, consumption probability is modelled and in the second step the consumed amount on consumption days is modelled (Tooze et al., 2006).

The currently published studies on two-part regression calibration method require calibration sub-studies with at least two replicates of reference measurement per subject (Kipnis et al., 2009; Tooze et al., 2006; Zhang et al., 2011). Given a single-replicate design of the EPIC study with zero-inflated reference measurements, however, the calibration models in the literature cannot be applied directly. Moreover, there is limited research on the performance of the two-part calibration model in a single-replicate study design for episodically consumed foods. Further, there is inadequate research on the effect of the standard theory of variable selection on the performance of a two-part calibration model in a single-replicate study design. The standard theory of selecting covariates into the calibration model states that confounding variables in the disease model must be included in the calibration model together with the covariates that only predict dietary intake (used as the response in the calibration model) but not the risk of the disease (Carroll et al., 2006; Kipnis et al., 2009).

To fill the aforementioned gaps, we developed a two-part regression calibration model to adjust for the bias in the diet-disease association caused by measurement error in episodically consumed foods, in the presence of a single-replicate calibration sub-study. The second goal was to assess the effect of reducing the number of variables selected into the two-part calibration model based on the standard theory. As a working example using the EPIC study, we studied the association between intakes of each of the three vegetable subgroups: leafy vegetables, fruiting vegetables, and root vegetables with all-cause mortality. We described how to handle the excess zeroes, skewness and heteroscedasticity in the reference measurements used as the response in the calibration model, nonlinearity, and how to select covariates into the calibration model. We showed that a suitably specified two-part calibration model adjusts for the bias in the diet-disease association caused by measurement error in self-reported intake. We further showed that the extent of adjusting for the bias is influenced by how the calibration model is specified, mainly with respect to forms of the continuous covariates.

2.2 Materials and Methods

2.2.1 Study subjects

EPIC is an on-going multicentre prospective cohort study to investigate the relation between diet and the risk of cancer and other chronic diseases. The study consisted of 519,978 eligible men and women aged between 35 and 70 years and recruited in 23 centres in 10 Western European countries (Riboli et al., 2002; Slimani et al., 2002). The 10 participating countries were: France, Italy, Spain, United Kingdom, Germany, The Netherlands, Greece, Sweden, Denmark, and Norway. The study populations comprised of heterogeneous groups. In most centres, study populations were based on general population while some consisted of participants in breast screening programs (Utrecht, The Netherlands; and Florence, Italy), teachers and school workers (France) or blood donors (certain Italian and Spanish centres). In Oxford, most of the cohort was recruited among subjects with interest in health or on vegetarian eating. Only women were recruited in France, Norway, Utrecht (The Netherlands) and Naples (Riboli et al., 2002). Information on usual dietary intake, lifestyle, environmental factors and anthropometry was collected from each individual at baseline. The dietary intake information was assessed with different dietary history questionnaires, food frequency questionnaires or a modified dietary history developed and validated separately in each participating country (Riboli et al., 2002). The questions asked in the questionnaires included the frequency of consumption over the past 12 months preceding the administration, categorized into the number of times per day, per week, per month or per year. A calibration sub-study was carried out within the entire EPIC cohort by taking a stratified random sample of 36,900 subjects. In the calibration sub-study, a 24-HDR was administered once per subject using a specifically developed software program (EPIC-SOFT) designed to harmonize the dietary measurements across study populations (Slimani, Valsta and Grp, 2002).

We used EPIC dietary intake data for leafy vegetables, fruiting vegetables and root vegetable sub-groups as a working example. We further assumed measurements from the 24-HDR (in g/day) as the reference measurements and the intake reported in the DQ as the main-study measurements. We excluded subjects with missing questionnaire data, missing dates of diagnosis or follow up, in the top and bottom 1% of the distribution of the ratio of reported total energy intake to energy requirement. We further excluded subjects with a history of cancer, myocardial infarction, stroke, angina, diabetes or a combination of these diseases at baseline. As a result, data for 430,215 subjects were eligible for the analyses. In the analysis, the data from the following centres were excluded: Umeå and Norway for leafy vegetables and Norway for fruiting vegetables. The decision to exclude these data was based on the inclusion criteria as stipulated in the EPIC analysis protocol.

2.2.2 Regression calibration model

In epidemiological studies, the interest is mainly in the association between an exposure and the risk of disease. In our working example, we were interested in the association between intake of vegetable subgroups and all-cause mortality. If the true usual intake of a vegetable subgroup is known, then the true association can be modelled by a generalized linear model (GLM) as

$$\varphi\{\mathrm{E}(Y \mid T, \mathbf{Z})\} = \beta_T T + \beta_{\mathbf{Z}}^T \mathbf{Z}, \tag{2.1}$$

where Y is a disease outcome, here, an indicator for mortality, T is true usual dietary intake of a vegetable subgroup, $\mathbf{Z}$ is a vector of error-free confounding variables and φ is a function linking the conditional mean and the linear predictor. The coefficient β_T quantifies the association of interest and $\beta_{\mathbf{Z}}^T$ is a vector of coefficients for the confounding variables. If dietary intake is measured with error, then β_T would mostly be underestimated.

Regression calibration is the most commonly used method to adjust for the bias in estimating β_T caused by measurement error in the DQ. To describe regression calibration, we denote reference measurement from the 24-HDR by R, main-study measurement from the DQ by Q, and the covariates that only predict vegetable intake and not all-cause mortality by $\boldsymbol{C}$. Therefore, a set of all covariates that possibly relate to usual intake is given by $\mathbf{X} = \{\mathbf{Z}, \mathbf{C}\}$. Regression calibration involves finding the best prediction of conditional expectation of true intake given DQ intake and other covariates assumed to be measured without error (Kipnis et al., 2009). The conditional expectation from regression calibration is denoted by $\mathrm{E}(T \mid Q, \mathbf{X})$. A major challenge in fitting the calibration model is that true usual intake is unobservable and cannot be measured exactly. As a result, a reference measurement is used in place of the latent true intake in the calibration model. Measurement from a valid reference instrument should be unbiased for true intake, and should have random errors that are uncorrelated with the measurement errors in the DQ (Kipnis et al., 2003). We, therefore, made two strong assumptions: that the 24-HDR is unbiased for true usual intake and measurement error in the 24-HDR is uncorrelated with the measurement error in the DQ. We denote the calibration model by:

$$\mathrm{E}(T \mid Q, \mathbf{X}) = \mathrm{E}(R \mid Q, \mathbf{X}) \tag{2.2}$$

We assumed in model (2.2) that measurement error in Q does not provide extra information about Y other than that provided by T. The measurement error in Q is, therefore, said to be non-differential. In model (2.2), R is modelled as a function of Q and $\boldsymbol{X}$ using standard regression methods, where a suitable distribution for the error terms and a suitable parametric form of each covariate in $\boldsymbol{X}$ is chosen.

In this work, we considered only the case of a single dietary intake variable measured with error. In our data, the correlation between the vegetable subgroups and the confounders, as measured by the questionnaire, were low justifying their omission, as the contamination effect of the measurement error in these variables on the correction factor for our dietary intake of interest would be negligible.

2.2.2.1 Excess zeroes, heteroscedasticity and skewness in reference measurements

Vegetable subgroups considered in this study are not consumed daily. This results in many zero reference measurements reported on the 24-HDR. As a result, the reference measurements have a mixture of zeroes for non-consumers and positive intake for consumers. To handle these excess zeroes, we used a two-part approach to build a regression calibration model. In the first part, the probability of reporting consumption in the 24-HDR is modelled. In the second part, the consumed amount given consumption in the 24-HDR is modelled (Tooze et al., 2006). The first part involves discrete data and can be modelled either with logistic or probit regression, where the probability of consumption is modelled conditional on a given set of covariates. In the second part, the consumed amount given consumption can be modelled conditional on the covariates and by assuming a plausible family of densities (McCullagh and Nelder, 1989). The GLM model for the consumption probability (Part I) is parameterized as

$$\mathrm{P}(R > 0 \mid Q, \mathbf{X}) = \phi^{-1}(\alpha_q Q + \alpha_{\mathbf{X}}^T \mathbf{X}) = \pi_{Q,\mathbf{X}},$$

where ϕ^{-1} can be either inverse-logit or inverse-probit function. Similarly, the GLM model for the consumed amount (Part II) is parameterized as

$$\mathrm{E}(R\,|\,Q,\mathbf{X};R>0)=g^{-1}(\beta_q Q+\beta_{\mathrm{X}}^T\mathbf{X})=\mu_{Q,\mathrm{X}}\,,$$

where g^{-1} can be an inverse of any plausible link function. Thus, the calibration model (2.2), adapted to two-part form to handle the excess zeroes in the reference measurements used as the calibration response is parameterized as

$$\mathrm{E}(R\,|\,Q,\mathbf{X})=\phi^{-1}(\alpha_q Q+\alpha_{\mathrm{X}}^T\mathbf{X})\times g^{-1}(\beta_q Q+\beta_{\mathrm{X}}^T\mathbf{X})=\pi_{Q,\mathrm{X}}\mu_{Q,\mathrm{X}}\,.$$

The true usual intake can thus be predicted from this two-part calibration model. We denote the prediction from this two-part calibration model by

$$\hat{\mathrm{E}}(R\,|\,Q,\mathbf{X})=\hat{\pi}_{Q,\mathrm{X}}\hat{\mu}_{Q,\mathrm{X}}\,. \tag{2.3}$$

Another challenge is how to handle distribution for the consumed amount that is commonly right-skewed and with heteroscedastic variance. To handle heteroscedasticity, we applied a GLM approach, where the variance is linked to the mean as

$$\sigma^2(R\,|\,Q,\mathbf{X};R>0)=\psi\{\mathrm{E}(R\,|\,Q,\mathbf{X};R>0)\}\,,$$

where ψ is a function that links the conditional variance with the conditional mean of consumed amount, $\sigma^2(\cdot|\cdot)$ denotes the conditional variance, and $\mathrm{E}(\cdot|\cdot)$ denotes the conditional expectation (Manning, Basu and Mullahy, 2005). The advantage of using the GLM approach is that the consumed amount can be predicted directly without transforming the response values in Part II of the calibration model. To determine the optimal relation between the conditional variance and the conditional mean, the GLM model shown above is parameterized using a class of power-proportional variance functions as follows

$$\sigma^2(R\,|\,Q,\mathbf{X};R>0)=\kappa\{\mathrm{E}(R\,|\,Q,\mathbf{X};R>0)\}^{\lambda}\,,$$

where κ denotes the coefficient of variation, λ is a finite non-negative constant. This power variance function can be rewritten in a linear logarithmic form as

$$\sigma(R\,|\,Q,\mathbf{X};R>0)=\mathrm{a}+\mathrm{b}\log\{\mathrm{E}(R\,|\,Q,\mathbf{X};R>0)\}\,, \tag{2.4}$$

where $a=(\log\kappa)/2$ and $b=\lambda/2$. In model (2.4), a value of λ equals zero refers to a classical nonlinear regression with constant error variance, λ equals one refers to a

Poisson regression with the variance that is proportional to the mean, where $\kappa > 1$ indicates degree of over dispersion. Similarly, λ equals two with $\kappa > 0$ refers to a gamma model with the standard deviation that is proportional to the mean (Manning and Mullahy, 2001). To explore a suitable value for λ to identify the right GLM model, we plotted centre-specific log-transformed standard deviation versus centre-specific log-transformed mean, separately for each of the three vegetable subgroups reported in 24-HDR in the EPIC study. The value of λ is estimated as twice the slope of the fitted regression line (see model (2.4)). The GLM model considered here can accommodate family of densities with skewed (asymmetric) distributions. We chose to use graphical method to identify λ due to its simplicity as opposed to estimation methods such as the maximum likelihood (MLE).

2.2.2.2 Nonlinearity and variable transformation

The relation between dietary intake variables is often nonlinear. To explore the form of relation between consumption probability as reported on 24-HDR and usual intake as reported on DQ, we applied two techniques: the empirical logit plot, and the nonparametric generalized additive model (GAM). With the empirical logit technique, we categorized DQ intake, starting with the category of never-consumers followed by 10 g/ day intake intervals. In each category, we computed the logit of consumption as reported on the 24-HDR. The formula for the empirical logit transformation used is given by (Cox, 1970; McCullagh and Nelder, 1989)

$$\log\left(\frac{y_i + 0.5}{n_i - y_i + 0.5}\right), \tag{2.5}$$

where y_i is the number of individuals who reported consumption on the 24-HDR and n_i is the number of individuals in the i^{th} DQ-category. The addition of 0.5 to both the numerator and the denominator of the logit function serves to avoid indefinite empirical logit values when $y_i = n_i$ or $y_i = 0$, and this particular value minimizes the bias in estimating the log odds (McCullagh and Nelder, 1989). The estimated empirical logit (computed from the 24-HDR intake) is plotted against the mean intake in the respective DQ-category (computed from the DQ intake). We fitted a loess curve to the resulting scatterplots to have a visual inspection of the form of relation between the two variables (Weiss, 2006). We further made the empirical logit plots for each of the participating country in the EPIC study. With the GAM technique, we obtained an optimal smoothing splines for the relation between the consumption probability reported in 24-HDR and DQ intake based on generalized cross validation criterion (GCV) (Hastie and Tibshirani, 1999).

We fitted the GAM model for consumption probability, assuming a binomial distribution and a logit link function using the mcgv package in R (Wood, 2012). In the GAM model, we included confounding variables in the disease model (**Z**). We used the partial prediction plot from the smoothed DQ component to identify plausible forms of parametric transformations for the DQ (Cai, 2008). From the selected set of parametric transformations, Akaike Information Criterion (AIC) was used to identify the optimal parametric transformation. Similar to the consumption probability part, we used the GAM approach to explore the optimal form of the DQ intake in the consumed amount part of the calibration model.

2.2.2.3 Variables inclusion in the calibration model

The theory of regression calibration states that all confounding variables in the disease model must also be included in the calibration model in addition to the covariates that only predict the response in the calibration model (Kipnis et al., 2009). We used the same set of confounding variables in Agudo (2004) that studied the relation between intake of vegetables and mortality in the Spanish cohort of EPIC. The eight confounding variables were: BMI (kg/m^2), smoking status (never, former, current smoker), physical activity index (inactive, moderately inactive, moderately active, active), lifetime alcohol consumption (g/day), level of education (none, primary, technical, secondary, university), age at recruitment (years), total energy (kcal), and sex (male/female). The covariates that only predict intake as measured 24-HDR were selected based on their statistical significance in the calibration model (2.3). We included plausible two-way interaction terms of DQ intake variable with the other covariates in the calibration model. We hereafter refer to each of the calibration model with covariates selected using the standard theory with the prefix “standard”, here, standard two-part calibration model. The covariates are included twice in the two-part calibration model (i.e., in each part of the two-part model), thus posing a threat to over fitting. Moreover, some disease confounding variables might not necessarily predict true usual intake. We therefore conducted a backward elimination on the standard two-part calibration model based on a significance level α of 0.2. We chose 0.2 to ensure that no significant covariates are excluded from the model. We hereafter refer to each of the reduced version of the standard calibration model with the prefix “reduced”, here, reduced two-part calibration model.

To assess the power of the probability part of the two-part calibration model to correctly discriminate consumers from non-consumers as reported in the 24-HDR, we used the Area under the curve from the Receiver operating characteristic curve of the fitted logistic model (Steyerberg, 2009). For the consumed amount part, we assessed the predictive power of the model based on the root mean squared error and the mean bias (Hastie, Tibshirani and Friedman, 2009). In building the two-part calibration model, we conducted country-specific rather than centre-specific regression calibration models to obtain stable estimates given the relatively smaller sample sizes in each centre (Ferrari et al., 2008). We also fitted other forms of regression calibration models to compare with the developed two-part calibration model. These forms of the other calibration model include

(i) A two-part calibration model similar to the developed one but with untransformed DQ. We hereafter refer to this model as “Two-part (untransformed DQ)”. The aim of fitting this model was to assess the effect of nonlinearity on the performance of a two-part calibration model.

(ii) A one-part calibration model with untransformed DQ and with the usual assumptions of a classical linear model. This is the calibration model commonly used by epidemiologists to adjust for the bias in the diet-disease association. In this model, two strong assumptions are made, namely, normality and linearity. The aim of fitting this calibration model was to quantify the inadequacy in adjusting for the bias in the diet-disease association when these assumptions are violated.

In each of these two forms of calibration models, we used the same set of covariates as in each part of the standard two-part calibration but with different parametric forms of the DQ intake as explained above. We conducted a backward elimination (α = 0.2) on each of these forms of regression calibration models to obtain their reduced forms. Subsequently, we used a Cox proportional hazard model in model (2.1) to study the association between usual intake of vegetable subgroups and all-cause mortality (Cox, 1972). The Cox proportional hazards model was stratified by centre and sex. To explore the form of relation between usual intake of each of the three vegetable subgroups and all-cause mortality in the Cox model, we plotted the log hazard ratio estimate against median intake in each DQ category (Sainani, 2009).

We used bootstrap procedure to compute correct standard error for the log hazard ratio estimate. The bootstrap approach accounts for the uncertainty in the calibration process. We used centre-stratified bootstrapping on the calibration sub-study. To each bootstrap sample, the main-study data was added and regression calibration model fitted to generate replicate versions of $\mathrm{E}(R \mid Q, \mathbf{X})$ for each subject in the entire EPIC cohort (Cassell, 2007). To each replicate data, the Cox model was fitted yielding an estimate of log hazard ratio with a standard error. The within-calibration and between-calibration variances were combined using Rubin’s formula to account for the uncertainty in the calibration process (Boshuizen et al., 2007; Geert Molenberghs, 2007; Rubin, 2004). The Rubin’s formula used to estimate the standard error for the log hazard ratio estimate is

$$\sigma^2{}_{\mathrm{T}}(\overline{\beta}) = \frac{1}{m}\sum_{i=1}^{m}\left(\mathrm{SE}(\hat{\beta}_i)\right)^2 + \left(1+\frac{1}{m}\right)\left(\frac{1}{m-1}\right)\sum_{i=1}^{m}\left(\hat{\beta}_i - \overline{\beta}\right)^2, \qquad (2.6)$$

where $\sigma^2{}_{\mathrm{T}}(\overline{\beta})$ is the total variance of the mean of log hazard ratio estimate from m calibrated samples, $\mathrm{SE}(\hat{\beta}_i)$ is the within-calibration standard error, and $\left(\frac{1}{m-1}\right)\sum_{i=1}^{m}\left(\hat{\beta}_i-\overline{\beta}\right)^2$ is the between-calibration variance.

We fitted a Cox proportional hazards model that ignores the measurement error in the DQ intake. This method is hereafter referred to as the naïve method. In the naïve method, the DQ intake measurements were used to study the association between usual intake of a vegetable subgroup and all-cause mortality.

2.3 Results

In Table 2.1, a high percentage of zeroes is shown in the 24-HDR intake for each of the three vegetable subgroups, especially for root vegetable subgroup in most of the participating countries. The rather high percentage of zeroes in the 24-HDR suggests that these subgroups of vegetables are not consumed daily by everyone. The Pearson correlation coefficient for each of the three vegetable subgroups in each of the participating countries, as measured with 24-HDR and DQ, were rather low but mostly statistically significant. The boxplots for the distribution of the consumed amount on consumption events as reported in the 24-HDR showed positive skewed distributions for these dietary intake variables (Figure 2.1). These exploratory findings suggested a need to properly handle the excess zeroes, to choose either a suitable distribution or a correct transformation for the consumed amount reported in the 24-HDR in building a calibration model.

Table 2.1: Country-specific summary measures for the percentage of zero intake measurements reported on 24-HDR (% R=0, non-consumers) and Pearson Correlation (ρ) for intake as measured by 24-HDR and DQ for leafy vegetables, fruiting vegetables and root vegetables. EPIC Study, 1999-2000

Participating Countries		Leafy vegetables		Fruiting vegetables		Root vegetables	
	N	% R=0	ρ	% R=0	ρ	% R=0	ρ
France	4735	42.8	0.17	44.4	0.10	71.6	0.06
Italy	3961	59.3	0.16	37.6	0.15	79.6	0.11
Spain	3220	48.9	0.34	31.7	0.22	76.1	0.12
UK	1313	68.2	0.16	40.8	0.19	59.3	0.23
Netherlands	4545	70.5	0.10	48.7	0.21	82.0	0.14
Greece	2930	67.9	0.10	29.5	0.13	83.2	0.03^{ns}
Germany	4418	75.9	0.15	41.6	0.17	79.2	0.22
Sweden	[a] 6132	70.5	0.19	34.9	0.24	67.2	0.17
Denmark	3918	77.4	0.09	41	0.21	61.8	0.40
Norway	[b]1798					58.5	0.12

[a]N is 3132 instead of 6132 for leafy vegetables in Sweden because data from Umeå were excluded from analysis based of the inclusion criteria in EPIC;

[b] N refers to data for root vegetables only because data for Norway were excluded for leafy vegetable and fruit vegetable subgroups;

[ns] means correlation is not statistically significant at $\alpha = 0.05$, other correlation coefficients are highly significant with $P < 0.0001$.

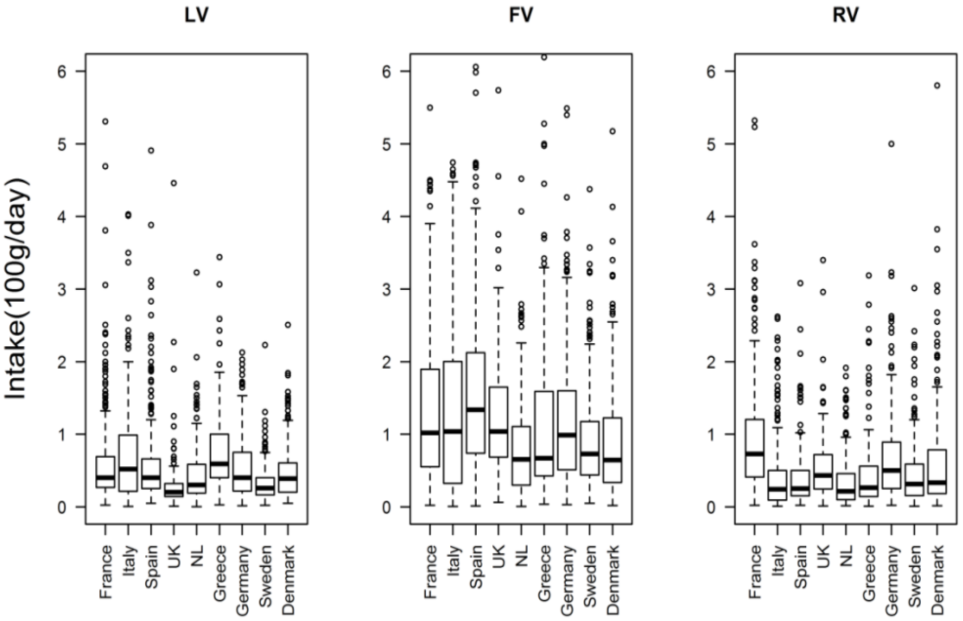

Figure 2.1: The country-specific boxplots show the distribution of the consumed amount for those who reported consumption on the 24-HDR for leafy vegetables (LV), fruiting vegetables (FV) and root vegetable (RV) subgroups in the EPIC study, 1992-2000.

For each of the three vegetable subgroups, a linear trend is shown between the log of standard deviation and the log of the mean for the consumed amount reported in the 24-HDR (Figure 2.2). The linear trend is a clear evidence of a variance that increases with a mean (presence of heteroscedasticity). The slope (standard error) of least squares regression line fitted to the resulting scatterplots was estimated as 1.057 (0.085). For fruiting vegetables, the estimates were 0.994 (0.076). Likewise for root vegetables, the estimates were 1.021 (0.130). These slopes of the fitted lines were all close to the theoretical value of 1 for a GLM gamma model. Based on these exploratory findings, we chose a gamma GLM model for the consumed amount in part II of the two-part calibration model for each of the three vegetable subgroups.

The correlation between each of the three vegetable subgroups ranged from 0.06 to 0.12 with total energy and from -0.07 to 0.05 with alcohol, as measured with the DQ.

These low correlations suggest minimal contamination effect of measurement error on diet-disease association, hence justifying our choice not to adjust for measurement error also for these confounding variables.

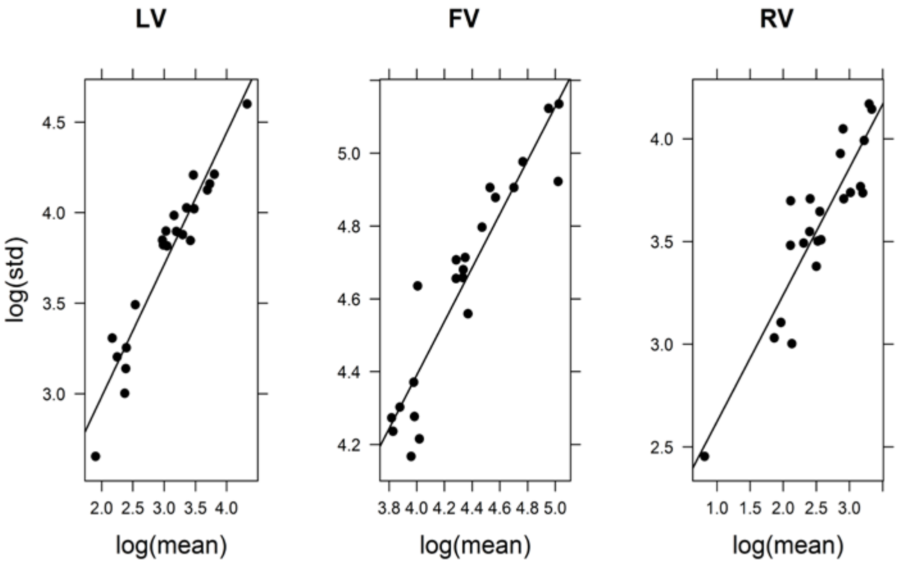

Figure 2.2: Variance-mean relation for leafy vegetable intake (LV), fruiting vegetable intake (FV) and root vegetable intake (RV). The graph shows a least squares regression line fitted to the scatterplots of the logarithm of centre-specific standard deviation versus logarithm of centre-specific mean of the consumed amount for those who reported consumption on the 24HDR in the EPIC Study, 1992-2000. The approximately linear regression line suggests a variance that increases with the mean.

To explore the form of the DQ intake variable for the consumption probability part of the two-part calibration model, the loess curve fitted to the scatterplots of the empirical logit versus the mean intake in each DQ category showed a nonlinear relation between the logit of consumption as reported in the 24-HDR and the DQ intake (dotted lines in Figure 2.3). The GAM partial prediction plots showed similar behaviour. From the plausible set of parametric transformations for the DQ, here, square-root and logarithmic, we chose log-transformed DQ based on the AIC criterion for each model fitted to country-specific data. As a result, we fitted a logistic model with log-transformed DQ and computed mean of the predicted logit of consumption in each category of the DQ. The loess curve fitted to the scatterplots of the mean predicted logit against the mean intake in a given DQ category is shown in the same figure (continuous line). The similarity of the two loess curves suggested the aptness of log-transforming DQ intake in the consumption probability model of leafy vegetables (Part I). The graphs for fruiting vegetables and root vegetables yielded similar results.

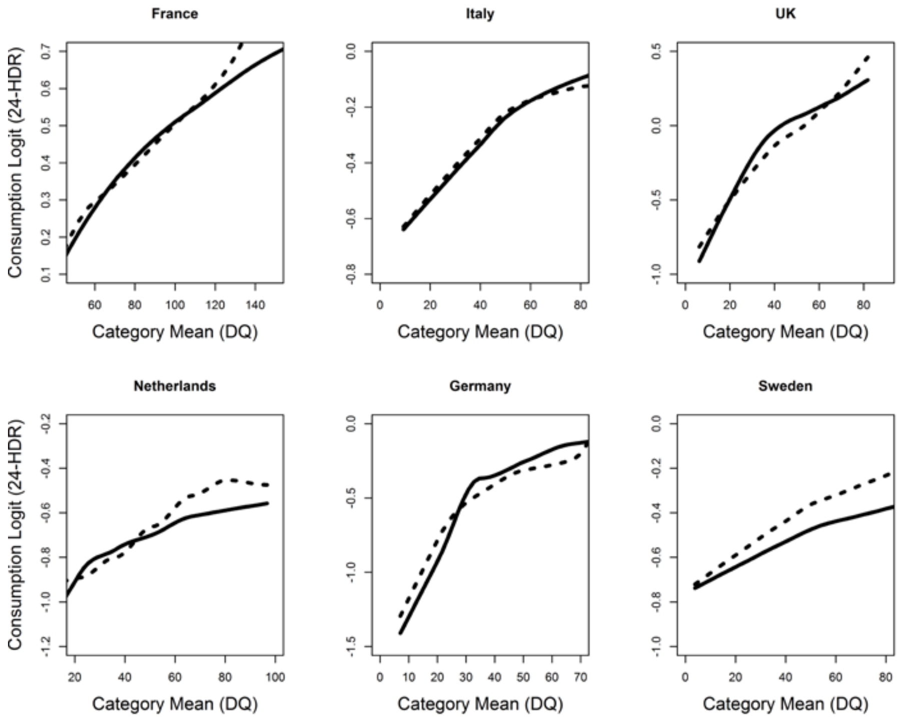

Figure 2.3: Country-specific empirical logit graphs for leafy vegetable intakes. The graph shows loess curves fitted to 1) the scatterplots for the empirical logit (dotted line) and 2) the mean of the predicted logit from a logistic model with log-transformed DQ (thick line) against the DQ category-specific means for leafy vegetable intake in the EPIC Study, 1992-2000. The similarity in the two logit curves suggests that a log- transformed DQ is appropriate for the consumption probability part of the two-part calibration model.

To explore the form of the DQ intake in the model for consumed amount (Part II), we fitted a GAM model with gamma distributed error terms and a log link function (as suggested by exploratory results). Based on partial prediction plots for the smoothed DQ intake component and using the AIC criterion, we chose a square-root transformed DQ intake for both leafy vegetables and root vegetables subgroups, and a log-transformed DQ intake for fruiting vegetables.

In addition to the confounding variables in the Cox proportional hazards model (shown in section (2.2.2.3)), season of DQ administration, centre where the DQ was administered and the body weight of the participant were also included in the calibration model, because they predicted intake of each of the three vegetable subgroups. Other covariates included in the standard two-part calibration model were the transformed DQ intake and its two-way with sex, age, season, BMI and centre. We used the same set of covariates on each part of the standard two-part calibration model, but with additional quadratic term for age at recruitment in the consumed amount part of the model. In Table 2.2, we showed the remaining significant terms after a backward elimination on each part of the standard two-part calibration model separately for each of the three vegetable subgroups.

Table 2.2: Significant covariates (marked ×) in the reduced two-part calibration models, after a backward elimination on each part of the standard two-part regression calibration model with transformed DQ and with other covariates selected using the standard way of variable inclusion. EPIC Study, 1992-2000

Covariates	**Leafy vegetables**		**Fruiting vegetables**		**Root vegetable**	
	Part I	Part II	Part I	Part II	Part I	Part II
Main effects						
Q^t	×	×	×	×	×	×
BMI		×	×	×	×	×
Smoking status	×		×	×	×	×
Physical activity		×	×	×		×
Lifetime alcohol	×		×			
Education	×	×	×	×	×	
Age	×	×	×	×	×	×
Age^2				×		
Total energy intake			×	×		×
Weight			×	×		×
Center	×	×	×	×	×	×
Season		×	×	×	×	×
Sex	×	×		×	×	
Interaction terms						
Q^t * sex		×			×	
Q^t * age	×	×		×		×
Q^t * season			×		×	
Q^t * BMI				×	×	
Q^t * center	×	×	×	×	×	×

Q^t is a transformed DQ; Part I, refers to consumption probability part of the two-part calibration model; Part II, refers to consumed amount part of the two-part calibration model; *, refers to an interaction term.

The areas under the curve from the ROC curve for the consumption probability part of the standard two-part calibration model and its reduced form were quite similar for each of the vegetable subgroups (Table 2.3). This suggest that some confounding variables and other two-way interaction terms of DQ intake with other covariates in the standard model do not necessarily predict the consumption probability and therefore should not be included in the calibration model. A similar remark could be made for the consumed amount part of the model, based on the root mean squared error and the mean bias, which were quite similar.

Table 2.3: The area under the curve (AUC) from ROC curve for consumption probability (Part I), and root mean square error (RMSE) and mean bias for the consumed amount (Part II) of the standard and the reduced forms of two-part regression calibration models with transformed DQ

Vegetable Subgroups	**Part I**		**Part II**	
	Models	AUC	RMSE[a]	Mean Bias[b]
Leafy	Standard	0.6846	66.841	0.0223
	Reduced	0.6843	64.578	0.0019
Fruiting	Standard	0.6305	118.823	0.0446
	Reduced	0.6304	110.415	-0.0334
Root	Standard	0.6413	68.626	0.0895
	Reduced	0.6408	66.524	0.0883

[a] $\text{RMSE} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{R}_i - R_i\right)^2$; [b] $\text{mean\_bias} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{R}_i - R_i\right)$

Figure 2.4, we fitted the smoothed curve to the scatterplots of the log hazard ratio estimate of dietary intake on all-cause mortality versus the median DQ intake in each DQ category. The graphical exploration showed approximately linear relations of DQ intake to all-cause mortality for each of the three vegetable subgroups. We therefore assumed a linear term for the DQ intake in the three fitted Cox proportional hazards models.

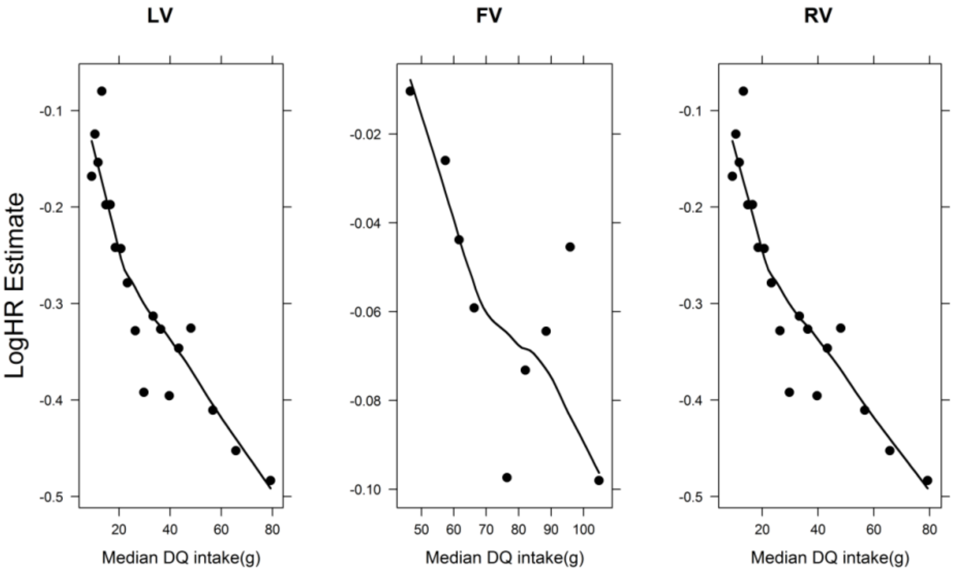

Figure 2.4: Assessment of linearity assumption of the DQ intake in the Cox proportional hazards model for leafy (LV), fruity (FV) and rooting (RV) vegetable subgroups. In each graph, a smoothed curve is fitted to the scatterplots of log hazard ratio estimate of vegetable subgroup intake on all-cause mortality versus DQ category-specific median intake (in grams). The approximately linear downward trend in each graph suggests an approximate linear relation and a beneficial effect of vegetable intake on the risk of all-cause mortality.

As expected, the log hazard ratio estimate for usual intake in the Cox model adjusted for measurement error in DQ intake were larger in absolute value than the naive estimate that ignores the measurement error. Similar remark was made for all the fitted forms of regression calibration models but the standard two-part calibration model with untransformed D (Table 2.4). The log hazard ratio estimates adjusted for the bias with the standard calibration models were smaller than those adjusted with the reduced calibration models, e.g., -0.265 for the standard two-part (transformed DQ) and -0.479 for the reduced two-part (transformed DQ) calibration model per 100g intake of root vegetables. The seemingly poor performance of the standard calibration models might be due to over fitting by covariates that did not significantly predict usual intake. The log hazard ratio estimate for root vegetables adjusted with the standard two-part calibration (with untransformed DQ intake) model ($\hat{\beta}_T$ = -0.107) was even smaller than the naïve estimate ($\hat{\beta}_T$ = -0.160). This suggests that a poorly specified functional form of a continuous covariate in a calibration model can result in adjusted association estimates that are more biased than the unadjusted estimates. The standard error of the log hazard ratio estimate, which is corrected for the uncertainty in the calibration, was larger than the uncorrected standard error for each of the calibration models. The underestimation of standard error was more severe for the standard calibration models. Further, the log hazard ratio estimate, which is calibrated with the reduced one-part linear calibration model, was smaller than that obtained with the reduced two-part (transformed DQ) model. Further, the predicted intake values for some subjects not in the calibration sub-study, in some cases were rather unrealistic. The unrealistic predictions were mainly from the standard calibration model with the untransformed DQ intake. The calibration models with the untransformed DQ intake resulted in a much smaller log hazard ratio estimate than their counterparts with the transformed DQ intake. This might be driven by extreme prediction from highly skewed DQ intake measurements in the calibration model. The effect of the extreme DQ values was further compounded by two factors: including the same covariate twice in the two-part calibration model and by the exponentiation effect due to the log link function used Part II of the calibration model. As a result, we assessed the sensitivity of the log hazard ratio estimate to these unreasonably high predicted values. Including these extreme predicted intake values resulted in massive change in the log hazard ratio estimate, mainly for the standard two-part calibration model with the untransformed DQ intake. For leafy vegetables, for instance, including the unrealistic predictions from the standard two-part calibration model with the untransformed DQ intake changed the estimate of log hazard ratio from -0.174 to -0.00518 per 100g intake. In Table 2.5 in the appendix, we present the percentages of these unrealistic predicted intake values, defined as extreme if it exceeded fivefold the ninety ninth percentile of the predicted usual intake. In the final analysis, we excluded these unrealistic values.

Table 2.4: Log hazard ratio estimate (standard error) per 100g usual intake of each of the three vegetable subgroups, calibrated with each of the three forms of regression calibration models in their reduced and standard forms

		Reduced model		Standard model	
Vegetable subgroups	Methods	$\hat{\beta}_T$ (SE^a; SE^b)	SE ratio^c	$\hat{\beta}_T$ (SE^a; SE^b)	SE ratio^c
Leafy	Naïve method	-0.144(0.027)		-0.144(0.027)	
	One-part linear calibration	-0.480(0.090;0.112)	1.24	-0.409(0.083;0.127)	1.53
	Two-part (untransformed DQ)	-0.395(0.092;0.183)	1.99	-0.174(0.089;0.278)	3.11
	Two-part (transformed DQ)	-0.509(0.090;0.292)	3.24	-0.461(0.047;0.160)	3.41
Fruiting	Naïve method	-0.094(0.014)		-0.094(0.014)	
	One-part linear calibration	-0.125(0.031;0.034)	1.11	-0.123(0.031;0.034)	1.11
	Two-part(untransformed DQ)	-0.161(0.030;0.034)	1.14	-0.109(0.030;0.073)	2.42
	Two-part (transformed DQ)	-0.255(0.037;0.108)	2.92	-0.228(0.035;0.131)	3.74
Root	Naïve method	-0.160(0.026)		-0.160(0.026)	
	One-part linear calibration	-0.342(0.060;0.082)	1.36	-0.305(0.054;0.077)	1.43
	Two-part(untransformed DQ)	-0.203(0.088;0.219)	2.49	-0.107(0.060;0.167)	2.78
	Two-part (transformed DQ)	-0.479(0.070;0.214)	3.06	-0.265(0.056;0.181)	3.23

SE^a is the standard error (×10^{-2}) for $\hat{\beta}_T$ that does not account for the uncertainty in the calibration; SE^b is the standard error (×10^{-2}) that accounts for the uncertainty in the calibration; SE ratio^c is the ratio of SE^b to SE^a

2.4 Discussion

In this work, we adapted a two-part regression calibration model initially developed for multiples of 24-HDR intake per individual for episodically consumed foods to a single replicate setting. We focused on dietary intakes with reference measurements that are skewed, heteroscedastic, and with substantial percentage of zeroes as reported on the 24-HDR. We further described how to explore and identify a suitable GLM model and a correct parametric form of a continuous covariate in the calibration model. As a result, we applied flexible GLM models that could simultaneously handle skewness and heteroscedasticity in the consumed amount. Thus, we avoided complications resulting from response transformation. We chose the log link function to stabilize the variance and to ensure positive prediction for usual intake (Raymond, 2010).

The standard way of including variables in the calibration model states that all confounding variables in the disease model and those that only predict dietary intake but not the disease occurrence must be included in the calibration model. Given the complexity of the two-part calibration model, some confounding variables in the disease model do not necessarily predict dietary intake. This could pose a threat to over fitting the calibration model. We conducted a backward elimination on each part of the two-part calibration model separately. The reduced calibration model with only significant covariates seemingly outperformed its standard counterpart in adjusting for the association bias. Leaving out confounding variables from the calibration model is against the standard theory of regression calibration. Nevertheless, we argue that if the omitted covariates have no effect in the calibration model, they should be excluded and the calibration method should still be correct. We further found that assuming linearity in the calibration model when it does not hold in a calibration model can pose a serious threat to the model performance. Similarly, Thoresen (2006) found in a simulation study that a less specified calibration model can have a considerable impact on the degree of bias-adjustment. We observed that predicted values for some subjects not in the calibration sub-study were extremely large. The extreme predictions resulted mainly from standard calibration models with linear term for the DQ intake. In such a case, predictions are made outside the variable space on which the model is fitted. As a consequence, the prediction space would extend more outside the variable space in the complex models.

The consumption probability and the consumed amount for episodically consumed foods may be correlated. In each of the fitted two-part calibration models, we accounted for this correlation partly by allowing covariates to overlap on both parts of the two-part calibration model (Kipnis et al., 2009). With only a single 24-HDR intake measurement per subject, part of correlation not explained by the covariates cannot be estimated. In future studies, a sensitivity analysis can be performed to assess the effect of the part of the correlation that is not explained by the covariates. This can be done by varying the magnitude of the assumed positive correlation between the consumption probability and the consumed amount in a simulation study (see chapter 3).

A limitation of this study is that we made some strong assumptions. First, we assumed the 24-HDR to be unbiased for true usual intake. Second, we assumed that the errors in the 24-HDR are uncorrelated with the errors in the DQ. However, previous studies have shown that these assumptions may not hold for dietary intakes, and that, the use of 24-HDR as a reference instrument for vegetable intake may be flawed (Kipnis et al., 2001; Kipnis et al., 2003; Natarajan et al., 2006; Natarajan et al., 2010). The biomarker studies using doubly labelled water for energy intake and urinary nitrogen for protein intake suggest that self-reports on recalls or food records may be biased. This is because individuals may systematically differ in their reporting accuracy. Additionally, the errors in these short-term instruments are shown to be positively correlated with the errors in the DQ (Day et al., 2001). As a result, using 24-HDR as a reference instrument can seriously underestimate true attenuation (Keogh, White and Rodwell, 2013). Therefore, the results obtained by using the 24-HDR as a reference instrument should be interpreted with caution. Nevertheless, the bias in the 24-HDR is reported to be substantially less severe than that in the DQ (Kipnis et al., 2001). Thus, when there is no objective biomarker measurements for a dietary intake, using the 24-HDR may still provide the best possible estimation of true intake (Kipnis et al., 2009).

In summary, a correctly specified two-part regression calibration model, which fits the data better, can adequately adjust for the bias in the diet-disease association, when only a single-replicate calibration sub-study is available. Further, the ability to adjust for the bias is influenced considerably by the form of the specified calibration model. We therefore advise researchers to pay special attention to calibration model specification, mainly with respect to the functional forms of the covariates.

Appendix

Table 2.5: Unrealistic predicted usual intake of vegetable subgroups. The maximum, the ninety-ninth percentile of predicted usual intake, and percentage (number) of unrealistic predictions using different forms of regression calibration models, each model in its reduced and standard forms

Vegetable subgroup	**Calibration method**	**Reduced form**			**Standard form**		
		Max	P99	% (n)	Max	P99	%(n)
Leafy	One-part [a]	224	76.8	0.00(0)	353	78.79	0.00(0)
	Two-part [b]	36292	133.7	0.00(12)	1.2×10^{10}	86.08	0.04(171)
	Two-part [c]	987.4	132.8	0.00(1)	13162.8	138.2	0.00(20)
Fruiting	One-part [a]	618	197	0.00(0)	591	200.7	0.00(0)
	Two-part [b]	3402	254.3	0.01(50)	3879	253.6	0.02(85)
	Two-part [c]	487.1	240	0.00(0)	511	207	0.00(0)
Root	One-part [a]	540.40[d]	65	0.00(0)	552.2	70.7	0.00(0)
	Two-part [b]	1.3×10^{10}	124.9	0.02(82)	2.3×10^{10}	132.2	0.02(89)
	Two-part [c]	2385	126.5	0.00(1)	27419.9	130.4	0.00(3)

Standard form: include all confounders in the Cox model and covariates that only predict intake; Reduced form: obtained by a backward elimination on the standard calibration model with α=0.2.Max: maximum predicted value; p99 is ninety ninth percentile of predicted intake.

[a] one-part linear calibration; [b] Two-part calibration with untransformed DQ; [c] Two-part calibration with transformed DQ; [d] the maximum value is not considered unrealistic despite being greater than fivefold ninety ninth percentile of predicted usual intake because it is within WHO recommendation of 400g daily intake.

3 Evaluation of a two-part regression calibration to adjust for dietary exposure measurement error in the Cox proportional hazards model: a simulation study [1]

Abstract

Dietary questionnaires are prone to measurement error, which bias the perceived association between dietary intake and risk of disease. Short term measurements are required to adjust for the bias in the association. For foods that are not consumed daily, the short term measurements are often characterized by excess zeroes. Via a simulation study, the performance of a two-part calibration model that was developed for a single-replicate study design was assessed by mimicking the European Prospective Investigation into Cancer and Nutrition (EPIC) study. The model was assessed with respect to the magnitude of the correlation between the consumption probability and the consumed amount (hereafter, cross-part correlation), the number and form of covariates in the calibration model, the percentage of zero response values, and the magnitude of the measurement error in the dietary intake. Transforming the dietary variable in the regression calibration to an appropriate scale was found to be the most important factor for the model performance. Reducing the number of covariates in the model could be beneficial, but was not critical in large-sample studies. The performance was remarkably robust when fitting a one-part rather than a two-part model. The model performance was minimally affected by the cross-part correlation.

[1] Based on: Agogo, G.O., van der Voet, H., van Eeuwijk, F.A. and Boshuizen, H.C. Evaluation of two-part regression calibration to adjust for dietary exposure measurement error in the Cox proportional hazards model: a simulation study: Biometrical journal (In press).

3.1 Introduction

Measurement error in exposure variables is a serious problem in relating health outcomes to exposures of interest (Carroll et al., 2012). In nutritional epidemiology, dietary exposures are usually measured with self-report instruments such as dietary questionnaires (hereafter, DQ). In the DQ, an individual is asked to recall his past intake over a long period of time, which might be difficult due to memory failure (Agudo, 2004). As a result, intake measurements from the DQ are prone to error. When DQ intakes are used to relate long-term intake with a health outcome, say, risk of a disease, the diet-disease association will be biased, typically toward a null value. Also, there will be loss of statistical power to detect significant associations (Kipnis et al., 2003).

To adjust for the bias in the association due to exposure measurement error, regression calibration is commonly used (Freedman et al., 2008; Rosner et al., 1989). Regression calibration involves finding the conditional mean predictor of true intake given DQ intake and other covariates, as explained in Carroll et al. (2006), p.66. Whereas DQ intake has measurement error, the other covariates are assumed to be measured without error. The conditional mean predicted intake is used instead of DQ intake in a model that relates the risk of a disease to dietary intake (hereafter, disease model). To apply regression calibration, a calibration sub-study is required, which contains unbiased measurements of true intake. This sub-study can comprise a random sample of the main-study population, where a short-term instrument, such as 24-hour dietary recalls (hereafter, 24HR), is used. When a dietary intake that is not consumed daily by everyone (hereafter, episodically consumed) is the exposure variable of interest, many zeroes will be recorded in the 24HR, denoting non-consumption. For such an episodically consumed food, the intake amounts on consumption days are usually right-skewed as shown in Kipnis et al. (2009), Rodrigues-Motta et al. (2015) and Tooze et al. (2006). The excess zeroes from the short-term instrument can be modelled using a two-part calibration model. In the first part, the mean probability of a non-zero response is modelled and in the second part, the conditional mean of a non-zero response value given consumption is modelled (Kipnis et al., 2009). Until now, the only methods that have been published use multiple replicate data.

Recently, Agogo et al. (2014) applied two-part regression calibration to adjust for measurement error in episodically consumed foods using calibration data with a single replicate of 24HR intake. In that paper, it was found that the performance of the calibration model can be influenced by the number and the form of the covariates included in the model. However, these findings were not investigated further. Regarding covariate selection, the standard theory states that confounding variables in the disease model must be included in the calibration model as described in Kipnis et al. (2009). Some of the confounding variables, however, may not significantly predict the response in the calibration model. Presently, there is limited research on the implication of the standard theory of variables selection in the two-part calibration model, especially for calibration studies with single replicate data. With a single-replicate calibration study, the magnitude of correlation between the probability of consumption and the consumed amount given consumption (hereafter, cross-part correlation) cannot be determined and is therefore assumed to be zero, given the covariates in the two-part model. It is hoped that allowing covariates to overlap on both parts of the two-part model can account for most of the cross-part correlation. This strong assumption requires proper investigation.

We assessed the performance of the two-part calibration model, when there is a calibration sub-study with single replicate data. Specifically, the following research questions were addressed:

(i) how the two-part calibration model performs as compared with a one-part calibration model,

(ii) how the two-part calibration model, with the covariates selected using the standard theory performs as compared with its reduced form that contains only the significant covariates,

(iii) how the two-part calibration model performs with varying magnitudes of the cross-part correlation,

(iv) how the two-part calibration model performs with varying percentages of zero values in the calibration response, magnitudes of measurement error in the DQ, and the strength of the association in the disease model, here, a Cox proportional hazards model.

The remainder of this paper is organized as follows: in section 3.2, the two-part regression calibration is described; in section 3.3, the simulation study is described. In section 3.4, different forms of calibration models are described and different measures to evaluate them are explained. The simulation results are shown in section 3.5. The study findings are discussed in section 3.6.

3.2 Regression calibration

The following notations are used: T_i for true long-term intake, Q_i for long-term intake from the DQ and R_i for unbiased intake measurement from the 24HR for the i-th individual. The vector of covariates measured without error is denoted by $\mathbf{Z}_i$, time to occurrence of disease by t_i. A possible association between dietary intake and time to occurrence of a disease can be modelled with a Cox proportional hazards model (Cox, 1972):

$$\mathrm{H}(t_i|T_i, \mathbf{Z}_i) = \mathrm{H_o}(t_i)\exp\left(\beta_1 T_i + \boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{Z}_i\right), \tag{3.1}$$

where $\mathrm{H_o}(\cdot)$ is the baseline hazard function, β_1 and $\boldsymbol{\beta}_2^{\mathrm{T}}$ are log hazard ratios (hereafter, logHRs). The interest is to estimate the logHR for intake (β_1). However, in practice, T_i is unknown and is often replaced by Q_i, leading to bias in the estimate of β_1. To adjust for this bias, regression calibration can be used, which involves finding the best conditional mean predictor of T_i given observed covariates: $\mathrm{E}(T_i|\,\mathbf{X}_i)$, where $\mathbf{X}_i = (Q_i, \mathbf{Z}_i^{\mathrm{T}}, \mathbf{C}_i^{\mathrm{T}})^{\mathrm{T}}$; $\mathbf{C}_i$ is a set of covariates that conditionally predict T_i given $\mathbf{Z}_i$ but do not predict t_i (Kipnis et al., 2009). As T_i is unobservable, the unbiased measurements R_i are used instead in a regression calibration to obtain

$$\mathrm{E}(T_i|\,\mathbf{X}_i) = \mathrm{E}(R_i|\,\mathbf{X}_i). \tag{3.2}$$

An important assumption in model (3.2), is that measurement error in Q_i is non-differential with respect to t_i (Keogh and White, 2014). This means that Q_i does not provide extra information about time to occurrence of disease t_i over what is contained in T_i (Carroll et al., 2006).

Because of excess zero R_i values for episodically consumed foods, model (3.2) can be partitioned into two parts. The first part models the mean probability of a non-zero R_i value (hereafter, consumption probability), $\mathrm{P}(R_i > 0| \mathbf{X}_i)$. The second part models the conditional mean of R_i given a non-zero R_i value (hereafter, consumed amount), $\mathrm{E}(R_i|\mathbf{X}_i; R_i > 0)$. Thus, long-term intake is the product of consumption probability and consumed amount:

$$\mathrm{E}(T_i| \mathbf{X}_i) = \mathrm{E}(R_i| \mathbf{X}_i) = \mathrm{P}(R_i > 0| \mathbf{X}_i)\mathrm{E}(R_i|\mathbf{X}_i; R_i > 0). \tag{3.3}$$

Model (3.3) can be fitted by assuming distributions such as the binomial for part I and gamma for part II as explained later in section 3.4.1. Such a procedure is referred to as two-part regression calibration. The same set of covariates can be used on both parts of the model. Typically, conditional independence given the covariates is assumed, when only a single-replicate data is available. When only fixed effects are present, model (3.3) assumes the same mean intake for two individuals with the same level of the covariates. However, an individual’s mean intake usually deviates from the population mean defined by $\mathbf{X}_i$ due to individual-specific random intake components. The random intake components on the two parts can be correlated. In order to estimate such a random effect model, one would need repeated measurements of R_i. The performance of model (3.3) for a single-replicate dietary data on simulated data with individual-specific random variations in true intake was assessed. Subsequently, the predicted T_i from the calibration model was used in the disease model (3.1) to estimate the association parameter β_1 reflecting the association of dietary intake with the time to occurrence of disease. The standard error of β_1 was estimated using a bootstrap method as explained later.

3.3 Simulation study set up

The simulation study was based on the design of the EPIC study. The EPIC study is an on-going multicentre prospective cohort study consisting of about half a million individuals aged mainly between 35 and 70 years, recruited in 23 centres in 10

European countries (Riboli et al., 2002; Slimani et al., 2002). We mimicked the distribution of leafy vegetable intake (hereafter, LV intake), which was the dietary variable of interest in this study. Leafy vegetables are not consumed daily, leading to many zeroes in the short-term instrument. In the EPIC study, long-term LV intake was measured with dietary food questionnaires specifically designed for each country or research centre within a country (Riboli et al., 2002). Within the EPIC cohort, an internal calibration sub-sample was randomly selected. From the calibration sub-sample, short-term intake measurements (in g/day) were recorded only once per individual using 24HR (Slimani et al., 2002).

In each of 16 simulated centres, observations for 2500 individuals were generated to reflect the large size of the EPIC study. Besides LV intake data, the other variables generated were: height (cm), weight (kg), BMI (kg/m2) computed from height and weight, gender, physical activity (inactive, moderately inactive, moderately active, active), highest level of education (none, primary, technical, secondary, university), age (years), total energy consumption (kcal), smoking status (never, former, current smoker), lifetime alcohol consumption (g/day), and season (autumn, winter, spring, summer) of administering the dietary questionnaire. To maintain the EPIC structure, these variables were generated by mimicking centre-specific correlation structures in the EPIC study (see Appendix 1).

3.3.1 Simulation model for true long-term intake ($\boldsymbol{T_i}$)

A logistic distribution was assumed for the true mean consumption probability of LV intake. The logit of consumption was assumed to be determined by age, BMI (Zhang et al., 2011), recruitment centre and gender. The centre effect reflects heterogeneity in consumption between centres. A lognormal distribution was assumed for the true intake amount, because many dietary intakes, including leafy vegetables are often skewed (Fraser and Stram, 2012). The consumed amount was also assumed to be determined by age, BMI, recruitment centre and gender. True long-term intake T_i, was generated as the product of the mean consumption probability π_{T_i} and the mean consumed amount on a consumption event A_{T_i}:

$$T_i = \pi_{T_i} A_{T_i}. \tag{3.4a}$$

In general, the consumption probability and the consumed amount given consumption can be correlated. The two components of true long-term intake T_i were simulated as follows

$$\pi_{T_i} = \text{expit}(\theta_{0g} + \theta_{1g} Age_i + \theta_{2g} BMI_i + \theta_{3g} centre_i \ + \sigma_{u_{\pi_T}} \epsilon_{1i}), \tag{3.4b}$$

$$A_{T_i} = \exp\{\lambda_{0g} + \lambda_{1g} age_i + \lambda_{2g} BMI_i + \lambda_{3g} centre_i + \rho\sigma_{u_{A_T}} \epsilon_{1i} + \sigma_{u_{A_T}} \epsilon_{2i} \sqrt{1-\rho^2}\}, \tag{3.4c}$$

where $\text{expit}(\cdot) = \exp(\cdot)/\{1+\exp(\cdot)\}$ is the inverse logit function, g is an index for gender (1 if male and 2 if female), θ_{0g} is the gender-specific intercept term, $\theta_{jg;j=1,2,3;g=1,2}$ and $\lambda_{jg;j=0,1,2,3;g=1,2}$ are gender-specific fixed covariate effect parameters, ϵ_{1i}, $\epsilon_{2i} \sim i.i.d.\,\text{N}(0,1)$ are standard normal random variables, ρ quantifies the correlation between the random intake components: $u_{\pi_{T_i}} = \sigma_{u_{\pi_T}} \epsilon_{1i}$ and $u_{A_{T_i}} = \rho\sigma_{u_{A_T}} \epsilon_{1i} + \sigma_{u_{A_T}} \epsilon_{2i} \sqrt{1-\rho^2}$. The random intake components are assumed to be independent of the covariates. If $\rho = 0$, then the two parts are correlated only through the overlapping covariates, else the two parts are correlated through both the overlapping covariates and the correlated random intake components. The aim was to assess how the magnitude of correlation in the random intake components affects the performance of the two-part calibration model. Therefore, the correlation in the random intake components ρ was set to 0 in most simulations, but was varied between -0.25 and 1.00 in a sensitivity analysis. Noteworthy, the simulation model assumes no never-consumers in the long run, because π_{T_i} and A_{T_i} are both nonzero.

The parameters used to generate T_i are presented in Table 3.1. The fixed effect parameters were obtained as the estimates from the EPIC data (Agogo et al., 2014). The intercept terms were adjusted in order to obtain the mean probability that is close to the mean frequency of a non-zero intake reported in the 24HR and the consumed amount that is close to the mean intake reported in the 24HR in the EPIC study. The variance components $\sigma^2_{u_{\pi_T}}$ and $\sigma^2_{u_{A_T}}$ were obtained as estimates from the Dutch National Food Consumption Survey of 2007-2010 (hereafter, DNFCS 2007-2010) as described in van Rossum (2011) and using the log-normal normal model in de Boer (2010).

Table 3.1: Parameter values for θ_{jg} and λ_{jg} fixed effect parameters for intercept, age, BMI and centre, and variances of random components used to simulate true long-term intake of leafy vegetables (refer to models (3.4b) and (3.4c).

Gender	model	Intercept	Age	BMI	Centre	Variance
Male	logit(π_{T_i})	-0.30	0.003	-0.01	0.012	0.09
	log(A_{T_i})	1.50	0.030	0.040	-0.003	0.25
Female	logit(π_{T_i})	-0.30	-0.002	-0.014	0.013	0.09
	log(A_{T_i})	1.50	0.030	0.030	-0.002	0.25

3.3.2 Simulation model for the unbiased measurement (R_i)

A calibration sub-study was randomly generated from the main-study population, consisting of 35% of individuals in the main study. The unbiased R_i were generated as follows. First, to characterize the many zero R_i values for LV intake, a uniform random variable U_i between zero and one was generated. Further, a zero or a non-zero R_i was randomly assigned to each individual in the calibration study by comparing U_i with π_{T_i}. A multiplicative measurement error structure for a non-zero R_i was assumed, because measurement error in dietary intake often increases with intake amount (Carroll et al., 2006).

The unbiased R_i were generated as follows

$$R_i = \begin{cases} A_{T_i}\exp(\epsilon_{R_i}),\ \epsilon_{R_i} \sim \mathrm{N}\left(-\frac{1}{2}\sigma_R^2, \sigma_R^2\right), \text{if } U_i \leq \pi_{T_i} \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad , \text{ otherwise} \end{cases} \tag{3.5}$$

where U_i~uniform [0,1]; $\sigma_R^2 = 0.69$ (from fitting the log-normal normal model in de Boer (2010) to the DNFCS 2007-2010 data) quantifies the within-person variation in intake. Using these parameter values the percentage of zero R_i values was about 50%.

3.3.3 Simulation model for the DQ (Q_i)

Leafy vegetable intake in the DQ was generated by mimicking a quantitative DQ, where not only the average consumption probability but also the average consumed amount is recorded. To generate mean probability of consumption, a linear measurement error model was assumed on the logit scale. The measurement error model contains constant bias ($\vartheta_1 \neq 0$), proportional scaling bias ($\vartheta_2 \neq 1$) and a random error term ($\epsilon_{\pi_{FFQ_i}}$) as

$\pi_{FFQ_i} = \text{expit}\{\vartheta_1 + \vartheta_2 \text{logit}(\pi_{T_i}) + \epsilon_{FFQ_i}\}, \ \epsilon_{\pi_{FFQ_i}} \sim \text{N}(0, \sigma^2_{\pi_{FFQ}})$,

where $\vartheta_1 = 3.6$ and $\vartheta_2 = 0.3$ (similar to estimates from the EPIC study).

Realistic questionnaires have a discrete format; therefore, the generated probabilities were translated according to a similar DQ categorization scheme as was used in the DNFCS 2007-2010. Following Goedhart (2012), specifically, the categorical answer for the mean probability of intake in the DQ was given as follows

1) $0 < \pi_{FFQ_i} <= 1/30$ 2) $1/30 < \pi_{FFQ_i} <= 3/30$ 3) $3/30 < \pi_{FFQ_i} <= 5/30$

4) $5/30 < \pi_{FFQ_i} <= 13/30$ 5) $13/30 < \pi_{FFQ_i} <= 21/30$ 6) $21/30 < \pi_{FFQ_i} <= 30/30$

The mid values, denoted by $FREQ_{FFQ_i}$, for the six categories are 0.017, 0.067, 0.133, 0.300, 0.567 and 0.850, respectively. Notably, if the generated probability π_{FFQ_i} falls within a particular category, the mid value of that category is used as the average consumption frequency in the DQ. There are two important sources of measurement error in the generated average consumption frequencies reported in the DQ. First, there is rounding error in the mid values that increases with the frequency of consumption. Second, there is potential misclassification in categorizing the probabilities into the above categories, due to measurement error in the simulated probabilities π_{FFQ_i}. These two sources of error are common in practice when using a questionnaire to measure dietary intake.

Similar to the true consumed amount, the consumed amount in the DQ was generated from a lognormal model, but with constant bias (κ_1), scaling bias (κ_2), centre-specific

bias (κ_3) and random error terms ($\epsilon_{A_{FFQ_i}}$). The centre-specific bias term κ_3 captures the effect of the centre on reporting of LV intake. The consumed amount in the DQ was generated as
$A_{FFQ_i} = \exp\{\kappa_1 + \kappa_2 \log(A_{T_i}) + \kappa_3 centre_i + \epsilon'_{A_{FFQ_i}}\}$,
$\epsilon'_{A_{FFQ_i}} = \log(\epsilon_{A_{FFQ_i}}) \sim \mathrm{N}(0, \sigma^2_{A_{FFQ}})$, where $\kappa_1 = -0.7$, $\kappa_2 = 1.5$ and $\kappa_3 = -0.1$ to mimic the systematic bias terms in the DQ estimates from the EPIC; $\sigma^2_{A_{FFQ}}$ is the random variation in reporting consumed amount. If the correlation between true intake and DQ intake is known, the error variance of DQ intake $\sigma^2_{A_{FFQ}}$ can be obtained using the correlation formula: $\rho_{T,Q} = \mathrm{Cov}(T,Q)/\sqrt{\mathrm{Var}(Q)\mathrm{Var}(T)}$, expressed in terms of parameters of the measurement error model similar to the formulation shown in Kipnis et al. (2003). The correlation between true and DQ intake is usually less than 0.7 (Fraser and Stram, 2001; Kaaks, 1997; Kaaks and Riboli, 1997). Similar to Fraser and Stram (2012), a modest correlation of 0.3 was assumed. The same approach was also used to estimate error variance in the frequency part $\sigma^2_{\pi_{FFQ}}$. The following variance components were obtained: $\sigma^2_{\pi_{FFQ}}$ as 0.57 and $\sigma^2_{A_{FFQ}}$ as 0.95. Long-term LV intake in the DQ was generated as

$$Q_i = FREQ_{FFQ_i} \times A_{FFQ_i}, \qquad (3.6)$$

the product of average consumption frequency and average consumed amount on a consumption day.

In order to assess the effect of the random measurement error in the DQ on the performance of the two-part calibration model, the magnitude of measurement error in Q_i was varied by multiplying $\sigma_{\pi_{FFQ}}$ and $\sigma_{A_{FFQ}}$ by a common factor denoted by λ.

3.3.4 Simulation model for survival time

For each individual, we let t_{D_i} be the time from study recruitment to all-cause mortality, and t_{E_i} be the time from study recruitment to the end of follow-up. Survival time is defined as the minimum of the two time intervals. Time to all-cause mortality, t_{D_i} was assumed to be exponentially distributed and to be determined by LV intake, age at recruitment, BMI, physical activity, alcohol consumption, cigarette smoking, total energy intake and education (Agudo et al., 2007). Following Bender, Augustin and Blettner (2005), t_{D_i} was generated as

$$t_{D_i} = -\frac{\log(U_i)}{\psi\exp(\beta_1 T_i + \boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{Z}_i)}, \tag{3.7}$$

where $\psi = 0.03$ is a constant baseline hazard, U_i~uniform [0,1]; β_1 and $\boldsymbol{\beta}_2^{\mathrm{T}}$ are logHRs. In line with the reported weak association between LV intake and all-cause mortality (Agudo et al., 2007), β_1 was taken as -0.005 (Agogo et al., 2014) and $\boldsymbol{\beta}_2^{\mathrm{T}}$= (0.01, 0.03, 0.2, -0.07, 0.5, 0.0001,-0.03). The time to end of study t_{E_i} was fixed as the 25th percentile of the time to all-cause mortality t_{D_i}, yielding 75% right censored observations to reflect low mortality rate usually observed in most epidemiologic studies.

3.4 Statistical analysis models

3.4.1 Standard two-part regression calibration

In the first part of the model, true mean consumption probability was predicted with a generalized linear model (GLM):$\mathrm{P}(R_i > 0|\mathbf{X}_i) = \Theta(\alpha_o + \boldsymbol{\alpha}_{\mathrm{z}}^{\mathrm{T}}\mathbf{X}_i)$, where $\Theta(\cdot)$ is an inverse logit link function and $\mathbf{X}_i$ is as defined in section 3.2. The $\mathbf{Z}_i$ component consisted of the same error-free covariates shown in model (3.7). The $\mathbf{C}_i$ component consisted of season of DQ administration, recruitment centre and gender. All the covariates in $\mathbf{X}_i$ except Q_i were assumed to be linearly related with the logit of consumption. To explore the form of relation and a suitable parametric transformation for Q_i that linearizes the relation, a single simulated dataset was analysed, using the empirical logit technique (Cox, 1970; McCullagh and Nelder, 1989). First, individuals were categorized based on their levels of intake as measured by Q_i. In each Q_i category, the logit of consumption was computed from R_i measurements. For each category, the logit of consumption was plotted against the median Q_i intake. Further, a smoothed curve was fitted to the scatterplots to explore the form of relation between Q_i and the logit of consumption. A transformation was chosen using the Akaike information criterion (AIC) to be either the logarithmic or the square root transform, which were a plausible set of simple transformations guided by the shape of the smoothed empirical logit curve described in detail in Agogo et al. (2014). In the second part, the mean consumed amount on consumption event was predicted from the following GLM model: $\mathrm{E}(R_i|\mathbf{X}_i; R_i > 0) = \Phi\{(\gamma_o + \boldsymbol{\gamma}_z^{\mathrm{T}}\mathbf{X}_i)\}$, where $\Phi(\cdot)$ is an inverse log link function that ensures positive predictions on the original scale. Further, a gamma distribution was assumed for the error structure. The strength of the gamma model is that it can handle skewness, heteroscedasticity and ensures predictions on the original scale. To explore an optimal transformation of Q_i that linearizes its relation with the consumed amount, a loess curve was fitted to the scatterplots of non-zero R_i against Q_i values for a single simulated dataset. As before, these curves were used to guide the choice of a suitable transformation. These transformations, chosen based on a single simulated dataset, were applied in all the simulated datasets.

With these two parts of the model, long-term dietary intake was predicted as

$$\widehat{\mathrm{E}}(R_i|\,\mathbf{X}_i) = \widehat{\mathrm{P}}(R_i > 0|\mathbf{X}_i)\,\widehat{\mathrm{E}}(R_i|\mathbf{X}_i;\ R_i > 0). \tag{3.8}$$

Model (3.8) is hereafter referred to as “Two-part RC (standard)”; “standard” here signifies that the vector of covariates $\mathbf{X}_i$ were selected using the standard theory of selecting all covariates that are in the disease model into the calibration model. Using the same covariates on both parts of the model can account for the cross-part correlation that is mainly caused by these covariates. Some covariates in $\mathbf{Z}_i$ that are included in $\mathbf{X}_i$, however, might not necessarily predict long-term intake and may result in over-fitting, especially given that they are included in the calibration model twice.

3.4.2 Reduced two-part regression calibration

Due to the potential threat to over-fitting, a backward elimination was performed on each part of the standard two-part calibration model separately based on a significance level (α) of 0.1. The selected set of significant covariates in each part of the model varied randomly from one simulation to the other. This model is hereafter referred to as “Two-part RC (reduced)”.

3.4.3 One-part regression calibration with transformed $\boldsymbol{Q_i}$

Unlike in the two-part calibration models, a one-part calibration model does not model the zeroes explicitly. In this model, R_i was assumed to be normally distributed. We used the same $\mathbf{X}_i$ and the same parametric form of Q_i as for the second (consumed amount) part of the standard two-part RC model. The aim of fitting this model was to assess the impact of misspecifying the calibration response in predicting intake. We hereafter, refer to this model as “One-part RC (transformed Q)”.

3.4.4 One-part calibration with untransformed $\boldsymbol{Q_i}$

This model is similar to the model in section 3.4.3 but with untransformed Q_i. The aim was to assess the effect of nonlinearity on the performance of the calibration model. We hereafter, refer to this model as “One-part RC (untransformed Q)”.

3.4.5 No calibration

For comparison, the Q_i measurements for the long-term LV intake were also used directly to relate intake with all-cause mortality. This method ignores measurement error in LV intake and is, hereafter, referred to as “Naïve method”. We also investigated the performance of the naive method for varying cross-part correlation, because it might be that the naïve method is sensitive to the magnitude of the cross-part correlation.

3.4.6 Disease model

The disease model (3.1), i.e., a Cox proportional hazards model was fitted using LV intakes, estimated as specified above. The model was adjusted for age at recruitment, BMI, physical activity, alcohol consumption, cigarette smoking, total energy intake and level of education. To account for the uncertainty in the calibration, the standard error of the logHR estimate was estimated using the bootstrap method (Efron and Tibshirani, 1993). In each simulation, a centre-stratified bootstrap sampling with replacement was performed to obtain 100 bootstrap samples. In each bootstrap sample, the logHR was estimated and the standard error estimate obtained as the standard deviation of the logHR estimates from 100 bootstrap samples.

3.4.7 Model evaluation

In this study, 1000 simulations were generated and analysed. The performance of each of the four calibration models: Two-part RC (standard), Two-part RC (reduced), One-part RC (transformed Q) and One-part RC (untransformed Q),was evaluated with respect to (a) the prediction of long-term intake and (b) the accuracy to estimate logHR of intake in the Cox model. The prediction of true long-term intake was assessed with

(i) Root Mean Square Error: $\mathrm{RMSE}(\hat{T}_i) = \frac{1}{\mathrm{B}}\sum_{\mathrm{k}=1}^{\mathrm{B}} \left[\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\hat{T}_i - T_i)^2\right]^{0.5}$, where n is the number of individuals, $\hat{T}_i$ is the predicted intake and B is the number of simulations,

(ii) Correlation coefficient between predicted and true intake.

The accuracy of the logHR estimate was assessed with (Burton et al., 2006):

(i) Absolute relative bias: $|\frac{(\bar{\hat{\beta}}_1 - \beta_1)}{\beta_1}|$, where $\bar{\hat{\beta}}_1$ is the mean logHR estimate from B simulations,

(ii) Root Mean Squared Error: $\mathrm{RMSE}\,(\hat{\beta}_1) = \sqrt{\left\{(\bar{\hat{\beta}}_1 - \beta_1)^2 + (\mathrm{Emp.SE}(\hat{\beta}_1))^2\right\}}$, where $\mathrm{Emp.SE}(\hat{\beta}_1)$ is the empirical standard error, defined as the standard deviation of $\hat{\beta}_1$ from B simulations,

(iii) Coverage probability: the proportion of times the 95% confidence interval of logHR estimate $\hat{\beta}_1 \pm 1.96\,\mathrm{SE}(\hat{\beta}_1)$ contains the true logHR,

(iv) Z score: $\bar{\hat{\beta}}_1/\overline{\mathrm{SE}}(\hat{\beta}_1)$, where $\overline{\mathrm{SE}}(\hat{\beta}_1) = 1/\mathrm{B}\sum_{i=1}^{\mathrm{B}}\mathrm{SE}(\hat{\beta}_{1i})$; z score is a measure of the statistical power to detect an association.

3.5 Simulation results

Figure 3.1 shows the histograms for the distribution of simulated true intake T_i (Figure 3.1a), unbiased measurement R_i (Figure 3.1b) and DQ intake Q_i (Figure 3.1c). The histograms show a skewed distribution of LV intake. The ‘spike’ at zero shown in the histogram of R_i indicates that LV intake is not consumed daily by everyone. Further shown are the loess curves for the scatterplots of the empirical logit computed from R_i versus median intake in Q_i category as measured with the DQ (Figure 3.1d), non-zero R_i versus Q_i (Figure 3.1e), and non-zero R_i versus log-transformed Q_i (Figure 3.1f). The loess curves suggest a non-linear relation between logit of a non-zero R_i and Q_i, and also between the non-zero R_i values and Q_i. In each of the two parts, a log-transformed Q_i seems to linearize the relation, for instance, as shown in Figure 3.1f for the relation between the simulated non-zero R_i values and Q_i. Based on AIC, in all cases logarithmic transformation of Q was used in both parts of two-part calibration model.

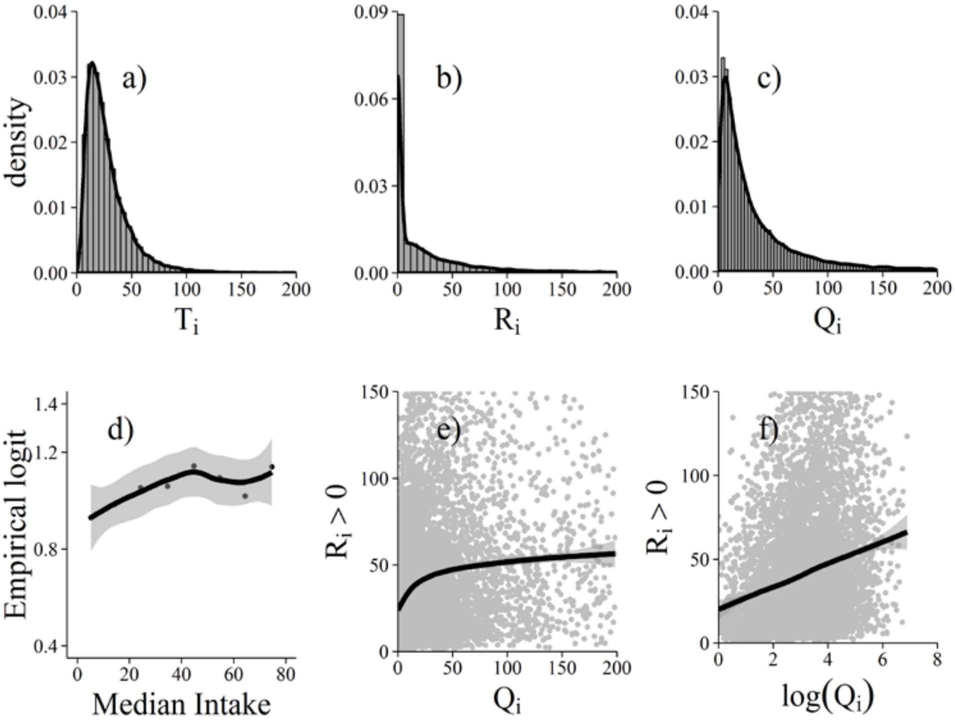

Figure 3.1: Histograms for the simulated true T_i (a), unbiased R_i (b) and DQ self-report Q_i (c) measurements, loess smoothed curve for empirical logit of a non-zero R value versus median intake level in Q_i categories (d), smoothed curve for the relation between non-zero R_i and Q_i (e) and smoothed curve for the relation between non-zero R_i and log transformed Q_i(f).

First, we compare the performance of two-part calibration (with transformed Q values), one-part calibration (either with or without transformation of the Q values) and the naïve method to the optimal method when true intake would be used. Table 3.2 presents the predictive measures for true intake and the respective logHR estimates. The power to predict intake declines for less specified calibration models. For instance, the one-part model (untransformed Q) predict true intake with the largest RMSE (17.34), whereas the two-part model (standard) predict true intake with the smallest RMSE (14.22). However, the one-part calibration model with transformed Q performs not much less than the two-part model (RMSE=14.90). Similarly, the accuracy and precision of the logHR estimate deteriorate for less specified calibration models; for instance, the RMSE of the logHR calibrated with the one-part model (untransformed Q) is 1.939 × 10^{3}, whereas the estimate calibrated with two-part model (standard) is 0.988 × 10^{3}. Again, the one-part model with transformed Q is close to the two-part model (RMSE=1.014 × 10^{-3}). Moreover, there is further loss of statistical power due to model misspecification as shown by smaller z score values for less specified calibration models. However, it is evident that regression calibration does not adjust for the loss of statistical power, as shown by the discrepancy between z score values for logHR estimate from the true regression and the estimates calibrated with regression calibration. Comparing the two forms of one-part calibration models, non-linearity markedly influence the calibration results; for instance, the standard error, empirical standard error and RMSE are almost doubled and the z score is almost halved when Q_i is included as a linear term in the one-part calibration model. As expected, the naïve method severely underestimates the true logHR and its standard error.

Table 3.2: Root mean squared error (RMSE) for predicted intake, correlation coefficient between predicted and true intake (ρ); mean, standard error, empirical standard error, RMSE, percent cover and z score for the log hazard ratio estimate for LV intake ($\hat{\beta}_1$), using three forms of regression calibration models, and naive estimate that ignores the measurement error

Method	RMSE ($\hat{T}_i$)	$\rho(\hat{T}_i, T_i)$	$\hat{\beta}_1$					
			$\bar{\hat{\beta}}_1$×10^{-2}	$\overline{\text{SE}}$×10^{-3}	Emp. SE×10^{-3}	RMSE ×10^{-3}	Percent Cover	z score
True regression	0.00	1.00	-0.500	0.590	0.586	0.586	95.30	-8.48
Two-part RC [a]	14.22	0.73	-0.487	0.998	0.979	0.988	93.10	-4.88
One-part RC [b]	14.90	0.70	-0.483	1.059	1.000	1.014	96.00	-4.56
One-part RC [c]	17.34	0.56	-0.476	2.029	1.924	1.939	95.20	-2.35
Naïve method		0.07	-0.004	0.057	0.060	4.957	0.00	-0.77

[a] two-part standard calibration with transformed Q;
[b] one-part calibration model with transformed Q;
[c] one-part calibration model with untransformed Q.

Next, the effect of having more zeroes in the calibration response was investigated. Table 3.3 displays the effect of varying the percentage of zero response values on the performance of the two calibration models: the standard two-part model and one-part model with transformed Q. The performance of the fitted calibration models declines with the increase in the percentage of zero response values as shown by larger RMSE for the predicted intake and larger absolute relative bias for the logHR estimate. The declining precision in prediction of the two-part calibration model with increasing percentage of zero response values is because the model for the intake amount is fitted to a small part of the data with nonzero response values, leading to large uncertainty in prediction, especially in extrapolating to the non-calibration sample.

Table 3.3: The overall means for true ($\bar{T}_i$), unbiased ($\bar{R}_i$) and predicted intake ($\bar{\hat{T}}_i$), the root mean squared error (RMSE) and Pearson correlation $\rho(\hat{T}_i, T_i)$ for calibrated intake and the estimated mean and absolute relative bias (Rel. Bias $\hat{\beta}_1$) for the log hazard ratio estimate ($\hat{\beta}_1$) adjusted for the bias with standard two-part RC model and one-part calibration model for various percentages of zero measurements in the calibration response

% R=0	Method	$\bar{T}_i$	$\bar{R}_i$	$\bar{\hat{T}}_i$	RMSE ($\hat{T}_i$)	$\rho(\hat{T}_i, T_i)$	$\bar{\hat{\beta}}_1\times$ 10^{-2}	Rel.Bias $\hat{\beta}_1$
10	Two-part RC [a]	27.15	27.15	27.14	12.72	0.74	-0.493	0.014
	One-part RC [b]	27.15	27.15	27.15	13.61	0.69	-0.499	0.002
50[c]	Two-part RC [a]	27.91	27.91	27.88	14.20	0.73	-0.488	0.024
	One-part RC [b]	27.92	27.89	27.89	14.90	0.70	-0.483	0.034
90	Two-part RC [a]	27.76	27.78	27.75	17.56	0.70	-0.454	0.092
	One-part RC [b]	27.77	27.73	27.74	17.80	0.69	-0.468	0.064

[a] two-part standard calibration with transformed Q;
[b] one-part calibration model with transformed Q;
[c] this is the percentage used also in the other simulations

In Table 3.4, the calibration results from the reduced two-part calibration model are compared with the results from the standard two-part calibration model. There is a slight improvement in the predictive power for true intake when the standard two-part calibration model is reduced. For instance, reducing the calibration model results in a smaller RMSE ($\hat{T}_i$) and a larger $\rho(\hat{T}_i, T_i)$ than the corresponding values from the standard two-part model. Additionally, the reduced calibration model leads to a logHR estimate with smaller uncertainty, larger coverage and larger z score values than the estimates from the standard calibration model. The improvement due to model reduction is, nevertheless, minimal because the standard calibration model is relatively simple with only the main effects of the covariates, and because of the large size of the dataset in the simulation study.

Table 3.4: Root mean squared error (RMSE) and Pearson correlation (ρ) for predicted intake ($\hat{T}_i$) and mean, standard error, RMSE, percent coverage and z score for log hazard ratio estimate ($\hat{\beta}_1$) adjusted for the bias with standard two-part RC and reduced two-part RC models (α= 0.1)

Method	RMSE $(\hat{T}_i)$	$\rho(\hat{T}_i, T_i)$	$\hat{\beta}_1$					
			$\bar{\hat{\beta}}_1$ $\times 10^{-2}$	$\overline{\text{SE}}$ $\times 10^{-3}$	Emp.SE $\times 10^{-3}$	RMSE $\times 10^{-3}$	Percent Cover	z-score
True regression	0.000	1.00	-0.500	0.590	0.586	0.586	95.30	-8.48
Two-part RC [a]	14.22	0.73	-0.487	0.998	0.979	0.988	93.10	-4.88
Two-part RC [b]	13.65	0.76	-0.488	0.992	0.928	0.935	95.80	-4.92

[a] two-part standard calibration with transformed Q; [b] reduced version of the two-part standard model

Finally, the effect of the cross-part correlation, i.e. correlation between consumption probability and consumed amount, was investigated, for the two-part calibration method and for the naive method. Table 3.5 presents the calibration results for various magnitudes of correlation between the two random intake components in the simulated long-term true intake. The correlation between consumption probability and the consumed amount increases with the increase in the correlation between the random intake components as expected. Even with zero correlation between the random intake components, a small positive correlation (0.05) is observed, which is induced by overlapping covariates on both parts of the two-part calibration model. Variability in true and predicted intakes and RMSE of predicted intakes increase as $\rho_{\pi_{T_i},A_{T_i}}$ increases. However, there seems to be a minimal effect of $\rho_{\pi_{T_i},A_{T_i}}$ on the calibrated logHR estimates. The results for the naïve method show that irrespective of the magnitude and sign of $\rho_{\pi_{T_i},A_{T_i}}$, the unadjusted logHR estimates are severely underestimated (not shown). Further, the relative bias for the unadjusted logHR estimate increases drastically for larger magnitudes of measurement error in the DQ, whereas the relative bias in the logHR estimate adjusted for the bias with the standard two-part regression calibration seems to be quite consistent, but with a slight increasing trend (Figure 3.2, Appendix 2). Also shown is an increasing trend in the relative bias for strong associations (Figure 3.3, Appendix 3).

Table 3.5:The nominal correlation coefficient between random intake components ($\rho_{u_{\pi_{T_i}},u_{A_{T_i}}}$), true ($\rho_{\pi_{T_i},A_{T_i}}$) and estimated ($\hat{\rho}_{\hat{\pi}_{T_i},\hat{A}_{T_i}}$) correlations between logit of consumption and log of consumed amount, standard deviations for true (T_i) and calibrated ($\hat{T}_i$) intake, and the root mean squared error (RMSE) for calibrated intake using standard two-part calibration model; the mean, standard error, Empirical standard error (Emp.SE) and RMSE of the log hazard ratio estimate ($\hat{\beta}_1$)

ρ	$\rho_{\pi_{T_i},A_{T_i}}$	$\hat{\rho}_{\hat{\pi}_{T_i},\hat{A}_{T_i}}$	Std T_i	Std $\hat{T}_i$	RMSE ($\hat{T}_i$)	$\bar{\hat{\beta}}_1 \times 10^{-2}$	$\overline{\text{SE}} \times 10^{-3}$	Emp.SE $\times 10^{-3}$	RMSE $\times 10^{-3}$	percent cover
-0.25	-0.06	0.01	19.58	14.78	12.98	-0.488	1.056	1.004	1.011	95.2
0.00	0.05	0.08	20.88	15.42	14.21	-0.489	0.998	0.965	0.971	93.1
0.25	0.17	0.14	22.12	16.09	15.34	-0.480	0.935	0.887	0.909	94.5
0.50	0.28	0.21	23.37	16.77	16.42	-0.485	0.887	0.839	0.852	94.8
0.75	0.39	0.26	24.57	17.46	17.43	-0.480	0.842	0.772	0.797	94.9
1.00	0.50	0.31	25.75	18.15	18.39	-0.480	0.803	0.763	0.788	94.9

ρ is nominal correlation coefficient between random intake components for consumption probability and consumption amount in the simulation model for long-term intake: $\rho_{u_{\pi_{T_i}},u_{A_{T_i}}}$.

3.6 Discussion

In this study, the performance of a two-part regression calibration was assessed in a validation study with single replicate data by mimicking a real epidemiologic study design. A gamma model was chosen in part II of the two-part calibration model, instead of the lognormal model used to simulate the data, thus mimicking the practical situation, where a statistical model, with a specified distribution, is assumed to approximate but not to exactly represent the unknown true distribution of the data.

The performance of the calibration model declined for less specified models. This was mainly due to an incorrectly specified functional form in the calibration model of the dietary questionnaire variable, which is a skewed continuous covariate, at least in this simulation study.

Notably, it was shown that reducing the two-part calibration model by applying variable selection criteria, did perform as well and could improve the performance of the calibration model. However, only a slight benefit of model reduction was observed. In principle, some covariates selected using the standard theory do not necessarily predict the calibration response and, therefore, could lead to overfitting the calibration model.

Moreover, in a two-part model, these covariates are included twice in the model (once in the probability part and once in the amount part), making the model even more vulnerable to over-fitting. However, our study used large sample sizes, leading to only minimal improvement due to model reduction. The benefit of a parsimonious calibration model would likely be larger for a study with a smaller sample size, for a study with highly zero inflated response values, or when a more complex calibration model is used. Even though this work showed some beneficial effect of reducing the calibration model, this might not be globally true and further work is still needed. For instance, it is worth exploring how the reduced model performs if the true long-term intake shown in equation (3.4a) is predicted by several weak predictors.

In the simulation study, the magnitude of correlation between random intake components did not substantially influence the log hazard ratio estimate. This could be due to our simulation set up, where a high correlation resulted in increased variability in the simulated true intake, which led to increased ability of detecting the effect of dietary intake on all-cause mortality.

The findings from this study are limited by the assumptions that were made. First, in our simulation we assumed a lognormal distribution for the consumed amount, whereas the true distribution for leafy vegetable intake, which is believed to be quite complex, can take another form. The assumed distribution, nevertheless, is a reasonable approximation due to its ability to capture common distributional features such as skewness for dietary intakes. Other distributions can be used such as a modified lognormal of Fraser and Stram (2012). Second, we focused on a univariate case where a single exposure variable is measured with error. Though not the focus of this study, in many epidemiologic studies, exposures are often measured with correlated errors (Marshall et al., 1999). The findings from this study, therefore, cannot be generalized for multiple exposures measured with correlated errors. The method described in this work, however, can be extended to a multivariate setting to adjust for measurement errors in multiple exposures (Fraser and Stram, 2001; Zhang et al., 2011). Further, instead of using a parametric transformation to model nonlinear relations, more robust semi/non-parametric methods could provide a better approximation of the underlying true relation (Hastie and Tibshirani, 1999). Likewise, instead of conducting a backward elimination to reduce variables in the calibration model, non-discrete methods, such as lasso, ridge, or elastic net, might provide better predictive calibration model (Hastie et al., 2009; Tibshirani, 1996, 1997). Lastly, even though in this study we simulated 24HR intake measurements as unbiased for true intake, in practice, the unbiasedness assumption for the 24HR might not hold, as previous studies showed 24HR to be marred by systematic bias in measuring protein and energy intake (Kipnis et al., 2003).

Notably, the true intake in our case refers to studies of chronic exposures. Studies of acute effects of short-term exposures will be more concerned with extreme intakes, rather than average intake. Our method therefore is not applicable to those situations.

The proposed model is closely connected to the models proposed by Tooze et al. (2006) and Kipnis et al. (2009), but differs from the two models in two important ways. First, we modelled the conditional mean amount (in part II) on the original scale using a GLM gamma model unlike a normal model on a Box-Cox transformed scale as used by the cited two papers. Second, in the proposed model, random effects are not included, because the method is applied to a calibration study with only a single replicate of a 24HR, unlike in the two papers with available multiple-replicate validation data. Beyond application in nutritional research, the proposed method can be applied to correct for covariate measurement error in other situations with zero-inflated data. Covariate measurement error has been found to be important in other areas such as mediation analysis (Zhao and Prentice, 2014), case-control studies (Guolo, 2008), environmental epidemiologic studies (Bateson and Wright, 2010) and in meta-analyses (Wood et al., 2009).

Appendices

Appendix 1. Simulating centre-specific correlation coefficients

First, a subset of EPIC data was created for individuals with complete information on height, weight, BMI (computed from height and weight), gender, physical activity, level of education, age, total energy intake, smoking status, lifetime alcohol consumption, season of administering the dietary questionnaire and study centre. Second, height, weight and total energy intake variables were log-transformed to make them more symmetrically distributed. Third, to maintain centre structure for the covariates, a centre-specific correlation matrix was estimated for each of the 16 centres using `corr` procedure in SAS. For each centre, the same variables were simulated with 2500 individuals to reflect the large size of the EPIC study. This was done using `simnormal` procedure in `SAS` version 9.3. Lastly, the variables that were simulated on a log-transformed scale were exponentiated.

Appendix 2. Magnitude of measurement error in the DQ

To assess the magnitude of measurement error in the dietary questionnaire (DQ) on the performance of the two-part calibration model, the random measurement error components on each part of the simulation model for DQ, shown by expression (3.6) in the main text, was multiplied by a common factor (denoted by λ). For each magnitude of λ, the log hazard ratio was estimated, after calibration with the standard two-part calibration model; also the estimate that ignored measurement error in the DQ was obtained. The results are shown below.

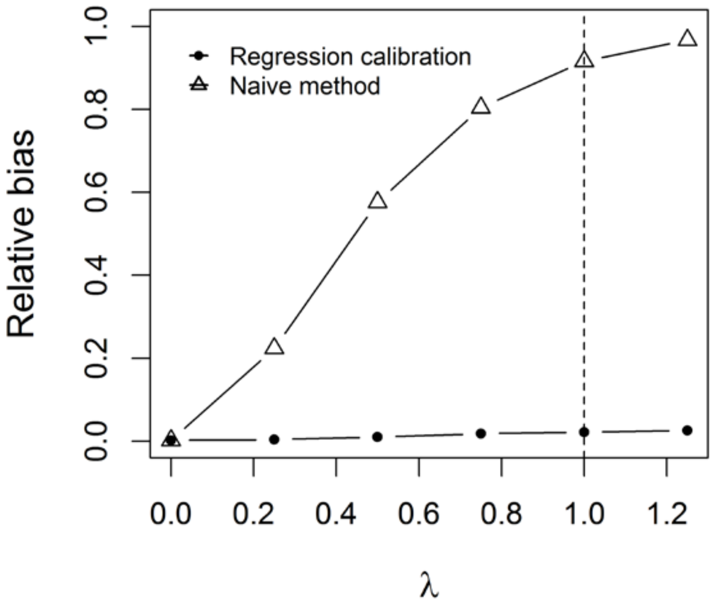

Figure 3.2: Absolute relative bias for log hazard ratio estimate $\hat{\beta}_1$ versus the multiplicative factor (λ) of the standard deviation of random error in each part of the simulation model for Q_i using the standard two-part calibration and the naïve method that ignores measurement error in Q_i. A larger λ value implies larger magnitude of measurement error. The dotted vertical line for λ =1 denotes the original simulation model.

Appendix 3. Strength of association in the Cox proportional hazards model

To assess the performance of the standard two-part calibration for varying strength of the diet-disease associations, the initial value for the log hazard ratio, i.e., -0.005, was multiplied by a factor of 2, 4, 6, 8 and 10 resulting in -0.01, -0.02,-0.03, -0.04 and -0.05, respectively. The absolute relative bias in the naïve analysis is consistently very high, whereas the bias in the calibrated log hazard ratio is remains consistently small, but with an increasing trend as shown below.

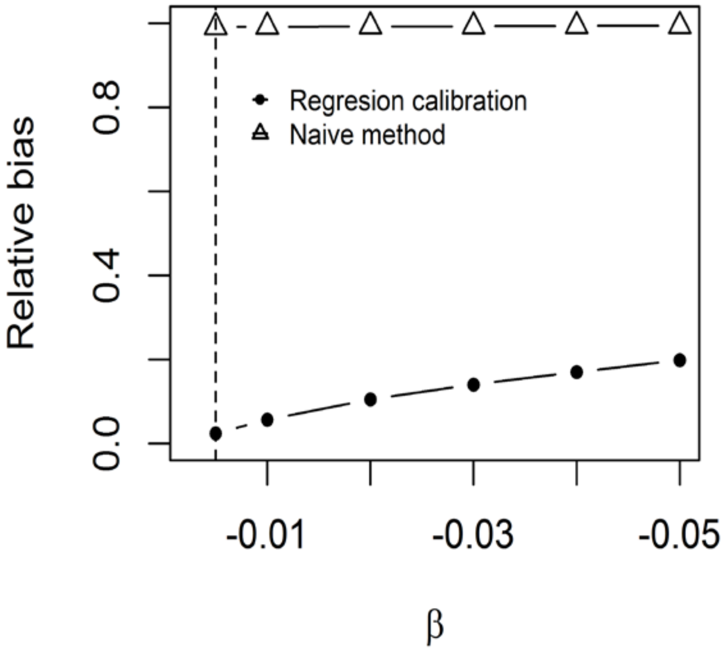

Figure 3.3: Absolute relative bias of the log hazard ratio estimate for in the Cox proportional hazard model using standard two-part regression calibration model by varying the magnitude of true log hazard ratio. The dotted vertical line denotes the standard two-part calibration model used in the main paper; the naïve results that ignore measurement error are also given.

4 A multivariate method to correct for Measurement Error in Exposure variables using External validation Data [1]

Abstract

Measurement error in self-reported dietary intake is known to bias the association between dietary intake and a health outcome of interest such as risk of a disease. The association can be distorted further by mismeasured confounders, leading to invalid results and conclusions. It is, however, difficult to adjust for the bias in the association when there is no internal validation data. We proposed a method to adjust for the bias in the diet-disease association (hereafter, association), due to measurement error in dietary intake and a mismeasured confounder, when there is no internal validation data. The method combines prior information on the validity of the self-report instrument with the observed data in order to adjust for the bias in the association. We compared the proposed method with the method that ignores the confounder, and the method that ignores measurement errors completely. We applied the methods to fruits and vegetables (FV) intakes, cigarette smoking (confounder) and all-cause mortality data from the European Prospective Investigation into Cancer and Nutrition study. Using the proposed method resulted in about four times increase in the strength of the association between FV intake and mortality. For weakly correlated errors, measurement error in the confounder minimally affected the hazard ratio estimate for FV intake. The effect was more pronounced for strong error correlations. The proposed method permits sensitivity analysis on measurement error structures and accounts for uncertainties in the reported validity coefficients. The method is useful in assessing the direction and quantifying the magnitude of bias in the association due to mismeasured confounders.

[1] Based on: Agogo, G.O., van der Voet, H., van Eeuwijk, F.A., van 't Veer, P., et al. Multivariate bias adjustment method to deal with measurement error in dietary intake using external validation data (submitted)

4.1 Introduction

The effect of measurement error on the association between an exposure and an outcome of interest has been studied extensively in epidemiology (Carroll et al., 2006; Day et al., 2004; Freedman et al., 2004; Freedman et al., 2008; Kipnis et al., 2003; Marshall et al., 1999), and particularly so in nutritional epidemiology. In nutritional research, the usually weak association between a dietary intake and the risk of a disease can further be distorted by another risk factor that is associated with both the disease and the dietary intake (hereafter, confounder) and by measurement error in the confounder. Moreover, the measurement error in the confounder can be more harmful in distorting the diet-disease association than the measurement error in the dietary intake (Marshall et al., 1999). If measurement error in the confounder is not taken into account, its effects can resonate so that a dietary intake with no effect can appear to have a sizable effect on the risk of a disease (Marshall et al., 1999). Resonant confounding can bias the diet-disease association in any direction, even when a researcher adjusts for confounding (Marshall et al., 1999; Wong, Day and Wareham, 1999). The resulting bias can be large.

In nutritional research, long-term dietary intakes are generally measured with dietary questionnaires (hereafter, DQs). The DQ is prone to recall bias that can result in either systematic bias or random error (Kipnis et al., 2003). The random error can be due to person-specific bias or within-person variation in intake. To validate the DQ, a validation study is required (Day et al., 2001; Natarajan et al., 2010). In a validation study, a short-term recall instrument or a biomarker is used to obtain unbiased measurements for an intake (hereafter, reference measurements) (Day et al., 2001; Subar et al., 2003). The reference measurements are used to quantify the effect of measurement error on the association parameter estimate. The effect of measurement error in the DQ can be quantified with either an attenuation factor or a correlation coefficient between true and measured intake (hereafter, validity coefficient)(Kipnis et al., 2003; Tooze et al., 2013). The attenuation factor quantifies the bias in the association estimate, whereas the validity coefficient quantifies the loss of statistical power to detect a significant association.

When only one risk factor is measured with error (hereafter, univariate case), a researcher can adjust for the bias in the association by dividing the unadjusted association estimate by the attenuation factor (hereafter, univariate method) (Rosner et al., 1990). However, complications may arise when confounders are also measured with error (hereafter, multivariate case) (Day et al., 2004; Wong et al., 1999). Measurement error in the confounder can contaminate the observed association. In the multivariate case, it is common for both dietary intake and confounder variables to be measured with correlated errors, further influencing the bias. Using the univariate method to adjust for the bias in the multivariate case can lead to substantial bias, especially for strong error correlations (Day et al., 2004). To adjust for the bias in the association using standard methods requires validation data from a validation study (Carroll et al., 2006; Freedman et al., 2011; Kipnis et al., 2015; Rosner et al., 1990). Generally, it is very costly to conduct such a validation study in addition to the main study.

We proposed a simple and flexible method to adjust for the bias in the diet-disease association caused by correlated measurement errors, in the absence of internal validation data. The proposed method demonstrates how to combine external data on the validity of the DQ with the observed DQ data to adjust for the bias in the diet-disease association and can be used to conduct sensitivity analysis on the effect of correlated measurement errors on study conclusions.

The method applies a Bayesian Markov Chain Monte Carlo (MCMC) sampling-based approach (Lesaffre and Lawson, 2012; Natarajan et al., 2010) and is implemented in SAS version 9.3. We illustrated the proposed method with data from the European Prospective Investigation into Cancer and Nutrition (EPIC) study. The aim in the EPIC example is to adjust for error in self-reported fruits and vegetables intake (hereafter, FV intake), when estimating the association of this dietary exposure with all-cause mortality, while simultaneously adjusting for the self-reported number of cigarettes smoked in a lifetime (hereafter, cigarette smoking), a variable believed to be also associated with all-cause mortality and also measured with error.

4.2 Materials and Methods

4.2.1 The EPIC Study example

The EPIC study is an on-going multicentre prospective study to investigate the association between nutrition and chronic diseases such as cancer (Riboli and Kaaks, 1997). In the EPIC cohort, baseline questionnaire and interview data on diet and non-dietary variables, anthropometric measurements and blood samples were collected. The study participants were followed over time for the occurrence of cancer, other diseases and overall mortality. The follow-up questionnaires were used to collect information on selected aspects of lifestyle that are related to the risk of cancer (Riboli et al., 2002). The EPIC study consisted of about half a million individuals aged mainly between 35 and 70 years, recruited in 23 centres in 10 European countries (Riboli and Kaaks, 1997; Slimani et al., 2002). Dietary food questionnaires (hereafter, DQs) were used to assess long-term dietary intake administered only once per subject. The mortality data were collected at the participating centres through mortality registries or follow-up and death-record collection (Riboli et al., 2002).

We used data for 9,341 individuals in the EPIC cohort who had complete information on FV intake, cigarette smoking, the study period and mortality. For illustration, we used FV intake as dietary intake, cigarette smoking as the confounder and whether a person died during the study period as an indicator of all-cause mortality. We illustrated the method with the aim of adjusting for the bias in the association between FV intake (in 100 grams per day) and all-cause mortality, while simultaneously adjusting for confounding by self-reported cigarette smoking and measurement error in cigarette smoking.

4.2.2 A measurement error model for the Dietary Questionnaire

We consider a Cox proportional hazards model to study the association between FV intake and all-cause mortality as

$$\mathrm{H}(t|T_1, T_2) = \mathrm{H_o}(\mathrm{t}) \exp(\beta_{T_1} T_1 + \beta_{T_2} T_2), \tag{4.1}$$

where $\mathrm{H_o(t)}$ is the baseline hazard at time to all-cause mortality t, β_{T_1} is the log hazard ratio (hereafter, logHR) for the true long-term FV intake T_1 and β_{T_2} is the logHR for the true confounder intake (cigarette smoking) T_2. For this study, the main interest is in estimating β_{T_1}. True FV intake, however, is unobservable in practice; therefore, the DQ measurement is usually used in place of the unknown true intake. Fitting model (4.1) to the observed DQ measurements for the FV intake (hereafter, Q_1) and cigarette smoking (hereafter, Q_2), replacing the corresponding true intakes, yields biased logHRs $\beta_{Q_1}$and β_{Q_2} of β_{T_1} and β_{T_2}, respectively. We refer to these biased log hazard ratios as unadjusted logHRs. Denote the vector of unadjusted logHRs $(\beta_{Q_1}, \beta_{Q_2})^{\mathrm{T}}$ by $\boldsymbol{\beta_Q}$ and a vector of true logHRs $(\beta_{T_1}, \beta_{T_2})^{\mathrm{T}}$ by $\boldsymbol{\beta_T}$. We assumed intake reported in the DQ to be linearly related to the true intakes, but with additional measurement errors (Keogh and White, 2014; Kipnis et al., 2003; Tooze et al., 2013) as

$$Q_i = \alpha_{0i} + \alpha_{1i} T_i + \epsilon_{Q_i}, i = 1{,}2, (1\text{= FV intake, 2= cigarette smoking}) \qquad (4.2)$$

where $(\epsilon_{Q_1}, \epsilon_{Q_2})^{\mathrm{T}} = \boldsymbol{\epsilon_Q} \sim \mathrm{N}(\boldsymbol{0}, \Sigma_{\epsilon_{\mathrm{Q}}})$, $(Q_1, Q_2)^{\mathrm{T}} = \boldsymbol{Q}$, the term α_{0i} quantifies the constant bias and α_{1i} quantifies intake-related/proportional scaling bias; the two components α_{0i} and α_{1i} jointly constitute systematic bias; the error component $\boldsymbol{\epsilon_Q}$ consists of person-specific bias and within-person random error terms and cannot be disentangled in a single-replicate study; ϵ_{Q_i} was assumed to be independent of true intake (T_i) and systematic bias components (α_{0i} and α_{1i}); person-specific bias component of $\boldsymbol{\epsilon_Q}$ describes the fact that two individuals who consume the same amount of FV will systematically report their intakes differently. Noteworthy, it is possible for the magnitude of self-reported intake to depend on the effects of subject's characteristics such as age and BMI. The contribution of these subject characteristic variables can be incorporated in model (4.2) by adding systematic terms attributable to these subject characteristic variables. Because the interest of this work was not in the effect of subject's characteristics on the validity of self-report instruments, we did not include their effects in the measurement error model. The unadjusted and true logHRs are linked as $\boldsymbol{\beta_Q} = \Lambda^{\mathrm{T}} \boldsymbol{\beta_T}$ (see supplementary information in Freedman et al. (2011)), where Λ is the attenuation-contamination matrix that quantifies the magnitude of attenuation, including contamination effects (the effects of error in measuring T_1 on β_{T_2} and the effect of error in measuring T_2 on β_{T_1}) (Rosner et al., 1990). The diagonal elements of Λ are the attenuation factors and the off-diagonal elements are the contamination factors (Freedman et al., 2011).

To adjust for the bias in the association using univariate method, a researcher simply divides each unadjusted logHR estimate of FV with the attenuation factor for the FV intake reported on the DQ (Freedman et al., 2011). Note that this method ignores the contamination effect of cigarette smoking.

To adjust for the bias in the association using the multivariate method, a researcher applies the inverse of the attenuation-contamination matrix to the unadjusted logHR as (Freedman et al., 2011; Rosner et al., 1990)

$$\hat{\beta}_{\boldsymbol{T}} = \left(\widehat{\Lambda}^{\mathrm{T}}\right)^{-1}\hat{\beta}_{\boldsymbol{Q}}, \tag{4.3}$$

where $\widehat{\Lambda}$ is usually estimated from a validation study. Many epidemiologic studies, however, do not conduct validation studies besides the main study, because validation studies are costly. We, therefore, proposed a method that incorporates external data on the validity of self-report instruments in estimating Λ. If no systematic bias is assumed in Q_i (i.e., $\alpha_{0i} = 0, \alpha_{1i} = 1$), Λ is the product of two covariance matrices: $\boldsymbol{\Sigma_T}$ for true intakes and $\boldsymbol{\Sigma_Q}^{-1}$ for the inverse of the covariance matrix of self-report intakes in the DQ and is estimated as $\widehat{\Lambda} = \hat{\boldsymbol{\Sigma}}_{\mathbf{T}}\hat{\boldsymbol{\Sigma}}_{\mathbf{Q}}^{-1}$ (see Carroll et al. (2006), p.362). With no systematic bias assumption, the elements required to obtain $\widehat{\Lambda}$ are:

$$\widehat{\Lambda} = \begin{pmatrix} \hat{\sigma}_{T_1}^2 & \hat{\sigma}_{T_1T_2} \\ \hat{\sigma}_{T_1T_2} & \hat{\sigma}_{T_2}^2 \end{pmatrix}\begin{pmatrix} \hat{\sigma}_{Q_1}^2 & \hat{\sigma}_{Q_1Q_2} \\ \hat{\sigma}_{Q_1Q_2} & \hat{\sigma}_{Q_2}^2 \end{pmatrix}^{-1}, \tag{4.4}$$

where $\hat{\sigma}_{T_1}^2$ and $\hat{\sigma}_{T_2}^2$ are variance estimates of T_1 and T_2, respectively. The covariance between true intakes is estimated as $\hat{\sigma}_{T_1T_2} = \hat{\rho}_{T_1T_2}\hat{\sigma}_{T_1}\hat{\sigma}_{T_2}$. Likewise, the covariance between the observed intakes reported in the DQ is estimated as

$$\hat{\sigma}_{Q_1Q_2} = \hat{\rho}_{T_1T_2}\hat{\sigma}_{T_1}\hat{\sigma}_{T_2} + \hat{\rho}_{\epsilon_{Q_1}\epsilon_{Q_2}}\hat{\sigma}_{\epsilon_{Q_1}}\hat{\sigma}_{\epsilon_{Q_2}}, \tag{4.5}$$

where $\hat{\rho}_{T_1T_2}$ is the estimate of correlation between true intakes and $\hat{\rho}_{\epsilon_{Q_1}\epsilon_{Q_2}}$ is the estimate of error correlation. Since $\hat{\Sigma}_{\mathbf{Q}}$ can be estimated directly from the observed DQ data, the task is to obtain $\hat{\sigma}^2_{T_1}, \hat{\sigma}^2_{T_2}$ and $\hat{\sigma}_{T_1T_2}$ in order to estimate all the elements in Λ shown in expression (4.4).

4.2.3 Estimation of $\Sigma_{\mathbf{T}}$ from DQ measurements and external validation data

We used the validity coefficient for the DQ to estimate the variance components of true intakes $\sigma^2_{T_1}$ and $\sigma^2_{T_2}$. Using parameters in model (4.2), the validity coefficient of the DQ is given by (Kipnis et al., 2003; Tooze et al., 2013)

$$\rho_{Q_iT_i} = \frac{cov(Q_iT_i)}{\sqrt{var(Q_i)var(T_i)}} = \frac{\alpha_{1i}\,\sigma_{T_i}}{\sigma_{Q_i}}.$$

From the validity coefficient formula, the variance for the true intake $\sigma^2_{T_i}$ can be estimated as

$$\hat{\sigma}^2_{T_i} = \left(\frac{\hat{\rho}_{Q_iT_i}}{\hat{\alpha}_{1i}}\,\hat{\sigma}_{Q_i}\right)^2, i = 1{,}2 \text{ (1 = FV intake, 2 = cigarette smoking).} \tag{4.6}$$

To obtain $\hat{\sigma}^2_{T_i}$, we need external data on the validity coefficient $\rho_{Q_iT_i}$ and the proportional scaling bias term α_{1i}. Hereafter, we set the proportional scaling bias to one ($\alpha_{1i} = 1$). The reason is that, at the time of this work, there were no previous studies with information on α_{1i} for FV intake and number of cigarettes smoked in a lifetime. However, this term can be incorporated in the measurement error model when dealing with study variables where this information is available. To obtain $\hat{\sigma}_{T_1T_2}$ from the observed data using expression (4.5), one has to make assumptions, as this information is generally not available from studies. The assumption can either be made on the correlation between true intakes $\hat{\rho}_{T_1T_2}$ or on the correlation between the errors $\hat{\rho}_{\epsilon_{Q_1}\epsilon_{Q_2}}$. The choice depends on the available prior knowledge for the study variables. The advantage of the proposed method is that it permits the user to make the assumption on either of the two correlations. A general assumption is that individuals often over report dietary intakes with health benefits, leading to positively correlated errors between variables with health benefits. Also, individuals often tend to under report intakes with harmful effects, leading to positively correlated errors between variables with harmful effects. Conversely, if one intake variable is over reported and the other under reported then one expects negatively correlated errors. We obtained a plausible range of validity coefficients from a literature review of studies on the validity of the questionnaire as a self-report instrument for long-term dietary intake T_1 and confounder intake T_2. We further assumed a plausible quantile of the uncertainty distribution for the range of validity coefficients ($\rho_{Q_iT_i}$) obtained from the literature. As no data are available for either of the two correlation coefficients $\rho_{\epsilon_{Q_1}\epsilon_{Q_2}}$or $\rho_{T_1T_2}$ for these two study variables, we assumed a range of possible values for these correlation coefficients. Noteworthy, by considering a range of values for the validity coefficient instead of just a single value, we accounted for the uncertainty due to heterogeneity between study populations in the literature reports.

4.2.4 A description of the proposed multivariate measurement error adjustment method

To adjust for the bias in the association parameters, we proposed a method that combines the observed self-report data in the DQ with the external validity data for the DQ derived from the literature. The method uses a Bayesian MCMC approach that accounts for the uncertainty in the literature reports, uncertainty that is both due to heterogeneity in the study populations in the literature reports and in the parameter estimation. Here, we described the bias-adjustment steps for the proposed method.

First, we obtained the posterior distributions of the unadjusted logHRs estimates $(\hat{\beta}_{Q_1}, \hat{\beta}_{Q_2})^{\mathrm{T}}$ that ignore the measurement error in the DQ. This was done by fitting a Bayesian Cox proportional hazards model (4.1) to the observed self-report data in the DQ for FV intake and cigarette smoking. In the Bayesian Cox model, we assumed weakly informative independent normal priors ($\pi_{\beta_{Q_i}}$) for the unadjusted logHRs by choosing a large variance as $\pi_{\beta_{Q_i}}$~$\mathrm{N}(0, 10^6)$.

Second, we estimated the posterior distribution of the covariance matrix for the observed self-report DQ data ($\Sigma_{\mathbf{Q}}$). Based on data exploration of the DQ data, a multi-normal distribution was assumed for the self-report intake data as $\mathbf{Q} \sim \mathrm{N}(\mu_{\mathbf{Q}}, \Sigma_{\mathbf{Q}})$. To ensure minimal influence of the prior information on the estimate of $\Sigma_{\mathbf{Q}}$, a weakly informative inverse Wishart prior ($\pi_{\Sigma_{\mathbf{Q}}}$) was assumed as $\pi_{\Sigma_{\mathbf{Q}}} \sim \mathrm{IW}(\Lambda_0, \upsilon_0)$, where $\Lambda_0 = \mathbf{I}_2$ (identity matrix) is the scale parameter and $\upsilon_0 = 2$ is the degrees of freedom. Note, this parameterization ensures a weak informative inverse Wishart prior for $\Sigma_{\mathbf{Q}}$ (Lesaffre and Lawson, 2012). Noteworthy, varying the magnitude of υ_0 did not alter the results much, because the likelihood dominated the prior given the large size of the EPIC data set.

Third, we generated the validity coefficients for the FV intake and cigarette smoking using prior information from the literature on external validation studies. We interpreted the lower and upper limits for the literature-reported validity coefficients as 0.05 and 0.95 quantiles of the distribution of plausible values, respectively. Using the limits of the literature-reported validity coefficients and the 90% quantile and because the distribution of correlation coefficients are usually skewed (Gorsuch, 2010; Lu, 2006), the validity coefficients were generated in a Fisher-z transformed scale as explained in Appendix 1. The generated validity coefficients were transformed back to the original scale using the inverse of Fisher-z transformation.

Fourth, using the validity coefficients generated from the literature data ($\rho_{Q_iT_i}$) and the posterior distribution for the variances of self-report intakes ($\sigma^2_{Q_i}$) estimated from the observed DQ data for FV intake and cigarette smoking, the corresponding distribution for the variance of true intakes ($\sigma^2_{T_i}$) was estimated as $\sigma^2_{T_i} = \left(\hat{\rho}_{Q_iT_i} \times \hat{\sigma}_{Q_i}\right)^2$ using expression (4.6), but with α_{1i} set to one.

Lastly, in order to estimate all the elements of Λ, we needed to estimate the covariance between true intakes $\hat{\sigma}_{T_1T_2}$. This could be done by decomposing the covariance in the observed DQ data $\hat{\sigma}_{Q_1Q_2}$ into the unknown covariance between true intakes $\hat{\sigma}_{T_1T_2}$ and the unknown covariance between the errors $\hat{\sigma}_{\epsilon_{Q_1}\epsilon_{Q_2}}$ as shown in expression (4.5). This covariance decomposition is only possible by making plausible prior assumption on either of the two covariances. Here, we made assumption on the plausible range of the covariance between the errors, because making this assumption is more intuitive for the two study variables in this work. To estimate the covariance between the errors $\hat{\sigma}_{\epsilon_{Q_1}\epsilon_{Q_2}}$, the error variance $\hat{\sigma}^2_{\epsilon_{Q_i}}$ was calculated as the difference between the estimated variance in the observed DQ data $\hat{\sigma}^2_{Q_i}$ and the estimated variance in true intake data $\hat{\sigma}^2_{T_i}$ as $\hat{\sigma}^2_{\epsilon_{Q_i}} = \hat{\sigma}^2_{Q_i}\,(1-\rho^2_{Q_iT_i})$. The remaining task is to estimate the unknown correlation between the errors ($\hat{\rho}_{\epsilon_{Q_1}\epsilon_{Q_2}}$) required to obtain $\hat{\sigma}_{\epsilon_{Q_1}\epsilon_{Q_2}}$. To our knowledge, there were no previous studies at the time of this work with information on the error correlation between FV intake and the number of cigarettes smoked in a lifetime. Due to lack of literature data on this error correlation, we generated the correlation between the errors $\rho_{\epsilon_{Q_1}\epsilon_{Q_2}}$ from a plausible range, guided by the correlation in the observed DQ data and the prior information on the most probable sign of the correlation between the errors in the FV intake and cigarette smoking (explained in the next section). With the generated $\rho_{\epsilon_{Q_1}\epsilon_{Q_2}}$, we could therefore obtain $\hat{\sigma}_{T_1T_2}$ as the difference between $\hat{\sigma}_{Q_1Q_2}$ and $\hat{\sigma}_{\epsilon_{Q_1}\epsilon_{Q_2}}$ parametrized as $\hat{\sigma}_{T_1T_2} = \hat{\sigma}_{Q_1Q_2} - \rho_{\epsilon_{Q_1}\epsilon_{Q_2}}\hat{\sigma}_{Q_1}\hat{\sigma}_{Q_2}\sqrt{\left(1-\rho^2_{Q_1T_1}\right)\left(1-\rho^2_{Q_2T_2}\right)}$. Thus, the distribution of the adjusted logHR for FV intake ($\hat{\beta}_{T_1}$) could be estimated from the distribution of $\left(\hat{\Lambda}^{\mathrm{T}}\right)^{-1}\hat{\beta}_{\boldsymbol{Q}}$ as shown in expression (4.3) and by following the above steps.

4.2.5 A comparison of the proposed method with other measurement error adjustment methods

We compared the results from the proposed multivariate method with (i) the results from applying the univariate method that ignores confounding by cigarette smoking and (ii) with the results from a method that ignores measurement error.

The proposed method was implemented in SAS version 9.3 using MCMC procedure as follows. The distribution of Fisher z-transformed validity coefficients were sampled directly from their prior distributions as explained above. The posterior distributions for the unadjusted logHRs estimates in the Bayesian Cox proportional hazard model were sampled using N-Metropolis method, with all initial parameter values set to zero. The convergence of the chains was assessed with the trace plots and autocorrelation with the autocorrelation plots. The analysis was based on 50 000 posterior samples, after discarding 5000 burn-in samples and using 5000 samples to tune the parameters. The results were summarized with density plots and posterior summary measures. We used R version 2.15.2 for graphing.

4.2.6 A sensitivity analysis

In our example, we investigated how different assumptions on the extent of measurement error in cigarette smoking affected the estimated logHR of FV intake $\hat{\beta}_{T_1}$. To do this, we used different values for the validity coefficients that were within the range reported in the literature. We further assessed how $\hat{\beta}_{T_1}$ varied with the magnitude of the correlation between the errors in FV intake and cigarette smoking. Lastly, we investigated the sensitivity of the results to the level of the uncertainty (expressed in quantile interval) assigned to the limits of the validity coefficients reported from the literature.

4.2.7 External data for FV intake and cigarette smoking

According to a pilot study on evaluation of dietary intake measurements in the EPIC study in nine European countries by Kaaks, Slimani and Riboli (1997) and a review study on FV intake by Agudo (2004), the validity coefficients of the DQ in measuring long-term FV intake is usually between 0.3 and 0.7. This range is consistent with the results reported from other similar validation studies (Feskanich et al., 1993; Goldbohm et al., 1994; SmithWarner et al., 1997). A validity coefficient greater or equal to 0.9 was considered as very uncommon (Agudo, 2004). According to Stram, Huberman and Wu (2002), the validity coefficient of self-reported cigarette smoking using cotinine as a marker for cigarette smoking, ranges mostly from 0.4 to 0.7. This range is consistent with the findings from other similar validation studies on adult smokers (Eliopoulos, Klein and Koren, 1996; Secker-Walker et al., 1997; Woodward, Moohan and Tunstall-Pedoe, 1999). A validity coefficient greater or equal to 0.85 was considered as very high (Stram, Huberman and Wu, 2002). Further, a weak negative correlation between beta-carotene (a marker for FV intake) and true lung dose was assumed, because smokers tend to have a poor diet (Shibata et al., 1992). Using these literature data to implement the proposed method, we interpreted the reported lower and upper limits of the validity coefficients as the 0.05 and 0.95 quantiles of the uncertainty distribution, respectively, to allow for all plausible values outside the reported range and to account for the population heterogeneity in these literature studies (see Figure 4.3 in Appendix 2).

Particular to FV intake and cigarette smoking, we assumed the error correlation to be mostly negative, because an individual will tend to over report his FV intake (a healthy habit) and to under report his cigarette smoking (an unhealthy habit). The assumed magnitude of error correlation, however, must be compatible with the correlation in the observed data. To ensure this compatibility, we obtained the upper limit of error correlation in the case that the correlation between true intakes is zero (i.e., the error covariance equals the covariance in the observed data) and assumed zero as the lower limit (i.e., the covariance in the observed data equals the covariance between true intakes). Similar to the validity coefficients, the lower and upper limits for the error correlation were assumed as 0.05 and 0.95 quantiles of the uncertainty distribution, respectively.

4.3 Results

Table 4.1 describes the logHR estimate for FV intake (per 100grams per day) and average number of cigarettes smoked per day, adjusted for the bias with the multivariate and the univariate methods; also shown are the unadjusted estimates. The logHR estimate adjusted for the bias with either the multivariate or the univariate method is greater in absolute value than the unadjusted estimate. The estimate adjusted for the bias with the multivariate method shows an about fourfold increase in the strength of association as compared with the unadjusted estimate. A similar magnitude of adjustment is shown with the univariate method. For cigarette smoking, both bias-adjustment methods give similar values for the logHR estimate. Further, the logHR for FV intake is estimated with a slightly larger uncertainty than the logHR for cigarette smoking. The similarity in the performance of the two bias-adjustment methods is due to the weak negative correlation between the errors assumed to be compatible with the correlation in the observed data (here, $\rho_{Q_1 Q_2} = -0.07$). The weak error correlation leads to a minimal contamination effect due to confounding by cigarette smoking. As expected, the variability in the unadjusted estimate is much smaller than the variability in the adjusted estimates for both intake variables. The small variability observed in the unadjusted estimates is because there is no uncertainty involved when measurement error is ignored in estimating the log hazard ratios.

Table 4.1: The mean (standard deviation), median, 0.05 and 0.95 quantiles, and mode for the Log Hazard Ratio (logHR) estimates for FV intake (per 100gram per day) and average number of cigarettes smoked (per day) adjusted for the bias with multivariate and univariate methods, and also the unadjusted estimates that ignore measurement error, EPIC study 1992-2000

	LogHR estimate for FV intake $\hat{\beta}_{T_1}$			
Methods	mean (SD)	median	90% CI	Mode
Multivariate	-0.181 (0.090)	-0.157	-0.375, -0.078	-0.125
Univariate	-0.169(0.082)	-0.147	-0.339,-0.077	-0.117
Unadjusted	-0.042 (0.007)	-0.042	-0.053,-0.031	-0.042
	LogHR estimate for cigarette smoking $\hat{\beta}_{T_2}$			
Methods	mean (SD)	median	90% CI	Mode
Multivariate	0.163 (0.079)	0.145	0.094;0.294	0.125
Univariate	0.162(0.077)	0.143	0.093;0.290	0.123
Unadjusted	0.046(0.002)	0.046	0.043;0.049	0.046

Abbreviation: CI is level of uncertainty in the range of literature-reported validity coefficient $\rho_{T_i Q_i}$ expressed as a credible interval.

Figure 4.1 displays the distribution for the estimates of the variance components required to estimate the attenuation-contamination matrix. The figure presents the kernel densities (curves) and means (solid vertical lines) of the variance estimates of the true intake levels and the mean estimate for the variance from the DQ measurements (dotted vertical lines) for FV intake (left panel) and cigarette smoking (right panel). From the graph, a large percentage of variability in the DQ is seemingly due to measurement error, and is influenced by the assumed magnitude of the validity coefficient. Based on this assumption, about 70% of variability in the DQ for both variables is due to measurement error. This means that only about 30% of the variability is attributable to inter-individual variability in true intake. The width of the density plot portrays the level of uncertainty involved in estimating the variance of true intake.

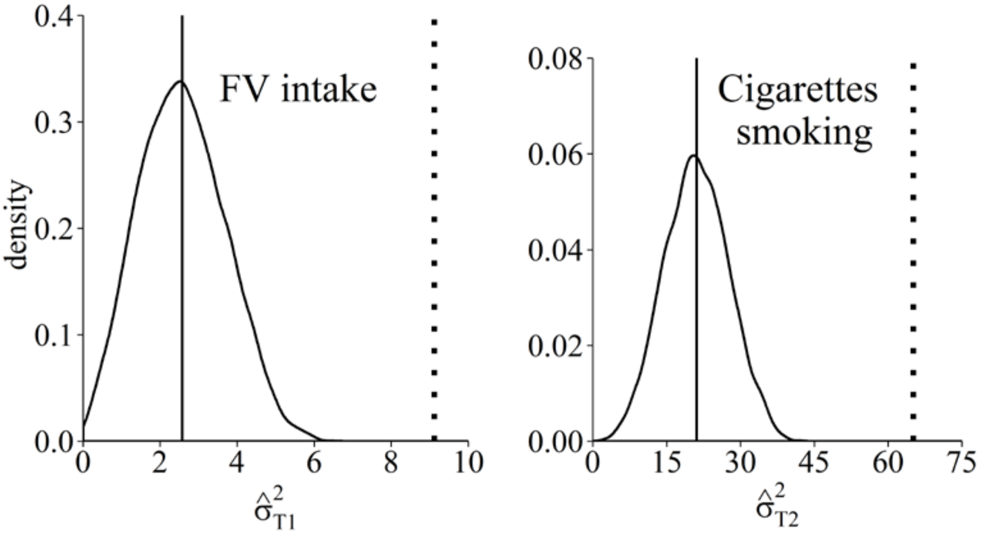

Figure 4.1: Kernel densities for the estimated posterior samples of variances for true intake levels of fruit and vegetable intake (FV intake, left panel) and true number of cigarettes smoked (right panel). The dotted vertical lines show the variance estimates from self-report in the DQ and the solid vertical lines show the posterior means of the estimated variances for true intake distributions.

Figure 4.2 shows the kernel densities and the means (solid vertical lines) for the estimates obtained with the multivariate method. The dotted vertical lines show the means of the unadjusted estimates. On average, the adjusted estimates are greater in absolute values than the unadjusted estimates, suggesting a stronger beneficial effect for FV intake (left panel) and stronger harmful effect of cigarette smoking (right panel). Importantly, in the multivariate case when both variables are measured with error, the unadjusted estimates can sometimes underestimate or overestimate the association, as hinted by the distribution of $\hat{\beta}_{T_1}$.The method estimates β_{T_1} with larger uncertainty (wider width) than β_{T_2}.

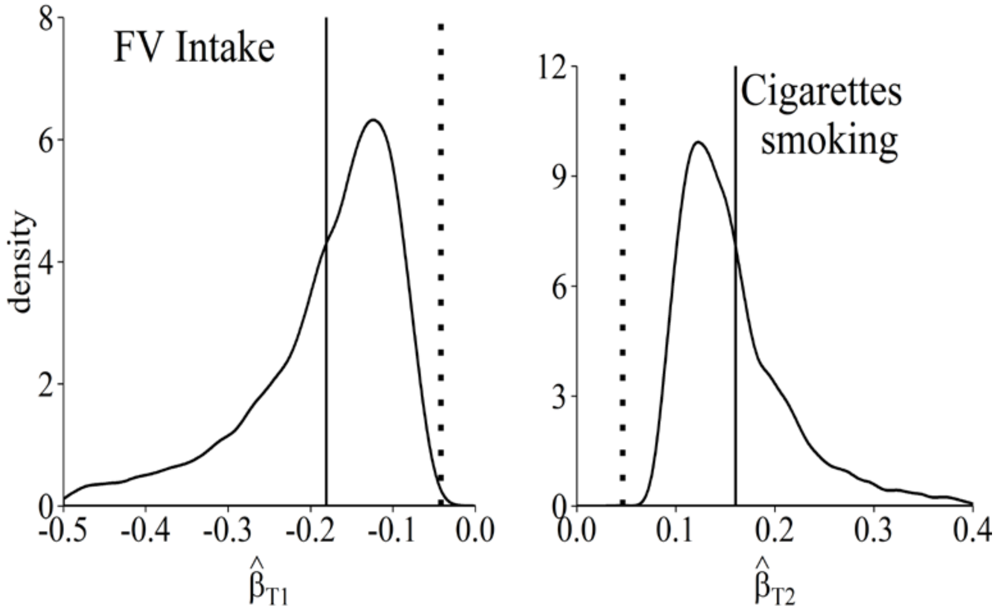

Figure 4.2: The kernel densities for the distribution of logHR estimates for fruits and vegetable intake per 100g per day ($\hat{\beta}_{T1}$, left panel) and for the number of cigarettes smoked per day ($\hat{\beta}_{T2}$, right panel) adjusted for the bias with the multivariate method. The dotted vertical line indicates the means of unadjusted logHR estimates; the solid vertical lines indicate the means of logHR estimates adjusted for the bias.

Table 4.2 presents the mean (standard deviation), median and mode of the logHR estimate for FV intake $\hat{\beta}_{T_1}$ and cigarette smoking $\hat{\beta}_{T_2}$ for various magnitudes of the validity coefficients of self-reported FV intake $\rho_{T_1Q_1}$ and self-reported cigarette smoking $\rho_{Q_2T_2}$. It is evident that the logHR estimate for FV intake $\hat{\beta}_{T_1}$ is influenced by the extent of measurement error assumed for cigarette smoking. For instance, when the validity coefficient for FV intake ($\rho_{Q_1T_1}$) is 0.5 and the validity coefficient for cigarette smoking ($\rho_{Q_2T_2}$) varies from 0.5 to 0.7, $\hat{\beta}_{T_1}$ is altered by about -3.8% (from -0.182 to -0.175). In contrast, the assumed magnitude of error in FV intake does not importantly influence the logHR estimate for the effect of cigarette smoking ($\hat{\beta}_{T_2}$), for instance, when $\rho_{Q_2T_2}$ is 0.5 and $\rho_{Q_1T_1}$ varies from 0.5 to 0.7, the value of $\hat{\beta}_{T_2}$ is almost the same. Noteworthy, if substantial error is assumed for cigarette smoking, $\hat{\beta}_{T_1}$ can become smaller than the unadjusted estimate, even when FV intake is assumed to be measured without error. The precision of the logHR estimates declines when larger measurement error is assumed for both variables. As expected, when both variables are assumed to be measured without error ($\rho_{T_1Q_1} = \rho_{T_2Q_2} = 1$), we get the same results as the unadjusted estimates.

Table 4.2: The mean (standard deviation), median and mode of log hazard ratio estimates for fruit and vegetables (FV) intake and number of cigarettes smoked adjusted for the bias with the multivariate method by varying magnitudes of validity coefficients assumed for the DQs for FV intake ($\rho_{T_1Q_1}$) and cigarette smoking ($\rho_{T_2Q_2}$), EPIC study 1992-2000

Validity coefficient		LogHR estimate for FV intake $\hat{\beta}_{T_1}$			LogHR for cigarette smoking $\hat{\beta}_{T_2}$		
$\rho_{T_1Q_1}$	$\rho_{T_2Q_2}$	mean (SD)	median	mode	mean (SD)	median	mode
0.3	0.3	-0.622 (0.227)	-0.605	-0.567	0.546 (0.048)	0.537	0.527
	0.5	-0.520 (0.101)	-0.517	-0.508	0.191 (0.012)	0.190	0.189
	0.7	-0.493 (0.083)	-0.491	-0.484	0.096 (0.005)	0.096	0.096
0.5	0.3	-0.207 (0.067)	-0.206	-0.203	0.522 (0.024)	0.522	0.521
	0.5	-0.182 (0.033)	-0.182	-0.181	0.187 (0.008)	0.187	0.187
	0.7	-0.175 (0.028)	-0.174	-0.173	0.095 (0.004)	0.095	0.095
0.7	0.3	-0.098 (0.030)	-0.097	-0.096	0.517 (0.021)	0.517	0.518
	0.5	-0.090 (0.017)	-0.090	-0.09	0.186 (0.008)	0.186	0.186
	0.7	-0.088 (0.015)	-0.088	-0.088	0.095 (0.004)	0.095	0.095
1.0	0.3	-0.029 (0.009)	-0.029	-0.029	0.513 (0.021)	0.514	0.517
	0.5	-0.038 (0.007)	-0.038	-0.038	0.185 (0.008)	0.185	0.185
	0.7	-0.041 (0.007)	-0.040	-0.040	0.094 (0.004)	0.094	0.095
	1.0	-0.042 (0.007)	-0.042	-0.042	0.046 (0.002)	0.046	0.046

Presented in Table 4.3 are the summary results for the logHR estimates adjusted for the bias with the proposed multivariate method by varying the assumed error correlation. It is evident that the magnitude of error correlation affects the mean estimate of the logHR for FV intake more than the mean estimate of the logHR for cigarette smoking. For positively correlated errors, the mean of $\hat{\beta}_{T_1}$ even becomes smaller in absolute value than the unadjusted estimate. Further, we compare the results obtained by assuming uncorrelated errors ($\rho_{\epsilon_1\epsilon_2} = 0$) in Table 4.3 with the results in Table 4.1. From this comparison, it is evident that the difference between the estimates obtained with the multivariate and univariate methods is due to the assumed magnitude of the correlation between true intakes ($\rho_{T_1T_2}$). When the errors are assumed to be uncorrelated, the presence of $\rho_{T_1T_2}$ alters $\hat{\beta}_{T_1}$ by about -6%, i.e., from -0.169 to -0.159 as estimated with the univariate method and the multivariate method, respectively.

Table 4.3: The mean (standard deviation), median, 0.05 and 0.95 quantiles and mode of the log hazard ratio estimates adjusted for the bias with the multivariate method by varying the magnitude of error correlation between DQ measurements for FV intake and number of cigarettes smoked, EPIC study 1992-2000

Correlations		LogHR estimate for FV intake $\hat{\beta}_{T_1}$			
$\rho_{\epsilon_1\epsilon_2}$	$\bar{\hat{\rho}}_{T_1T_2}$	mean (SD)	median	90% CI	Mode
-0.2	0.51	-0.301(0.098)	-0.294	-0.471, -0.155	-0.237
-0.2	0.38	-0.277(0.099)	-0.264	-0.460, -0.137	-0.212
-0.1	0.24	-0.247(0.098)	-0.228	-0.440, -0.117	-0.178
-0.1	0.10	-0.207(0.093)	-0.184	-0.403, -0.096	-0.143
0.0	-0.04	-0.159(0.083)	-0.136	-0.337, -0.069	-0.106
0.1	-0.32	-0.038(0.098)	-0.045	-0.171, 0.126	-0.047
Correlations		LogHR estimate for cigarette smoking $\hat{\beta}_{T_2}$			
$\rho_{\epsilon_1\epsilon_2}$	$\bar{\hat{\rho}}_{T_1T_2}$	mean (SD)	median	90% CI	Mode
-0.2	0.51	0.183(0.064)	0.169	0.109, 0.304	0.151
-0.2	0.38	0.178(0.067)	0.163	0.105, 0.305	0.143
-0.1	0.24	0.173(0.070)	0.156	0.101, 0.303	0.135
-0.1	0.10	0.167(0.075)	0.148	0.097, 0.295	0.130
0.0	-0.04	0.161(0.083)	0.141	0.093, 0.286	0.118
0.1	-0.32	0.157(0.075)	0.137	0.087, 0.294	0.116

CI is level of uncertainty in the range of literature-reported validity coefficient $\rho_{T_iQ_i}$ expressed as a credible interval; $\bar{\hat{\rho}}_{T_1T_2}$ is posterior mean estimate for the correlation coefficient between true intake variables.

Table 4.4 presents the mean (standard deviation), median, 0.05 and 0.95 quantiles and mode for logHR estimates $\hat{\beta}_{T_1}$ and $\hat{\beta}_{T_2}$ adjusted for the bias with the proposed multivariate method for various possibilities of equating the limits on literature-reported validity coefficients to quantiles of the uncertainty distribution. From this sensitivity result, the level of uncertainty assumed in the distribution of validity coefficient has negligible effect on the mean and the mode but not the median estimates of $\hat{\beta}_{T_1}$ and $\hat{\beta}_{T_2}$. As expected, the uncertainty in the estimates increases with the level of uncertainty assigned to the validity coefficients.

Table 4.4: The mean (standard deviation), median, 0.05 and 0.95 quantile and mode for logHR estimates for FV intake and for number of cigarettes smoked adjusted for the bias with the multivariate method, for various possibilities of equating the limits of literature-reported validity coefficients to quantiles of the uncertainty distribution, EPIC study 1992-2000

CI (%)	LogHR estimate for FV intake $\hat{\beta}_{T_1}$			
	mean (SD)	median	90%CI	Mode
80	-0.206 (0.155)	-0.156	-0.545, -0.072	-0.105
90	-0.181(0.090)	-0.157	-0.375, -0.078	-0.125
95	-0.179 (0.080)	-0.158	-0.348, -0.088	-0.155
99	-0.173 (0.065)	-0.160	-0.300, -0.095	-0.135
CI (%)	LogHR estimate for cigarette smoking $\hat{\beta}_{T_2}$			
	mean (SD)	median	90%CI	Mode
80	0.178 (0.128)	0.142	0.086, 0.381	0.142
90	0.163 (0.079)	0.145	0.094, 0.294	0.125
95	0.157 (0.056)	0.145	0.099, 0.257	0.122
99	0.150 (0.035)	0.144	0.107, 0.215	0.131

CI is level of uncertainty in the range of literature-reported validity coefficient $\rho_{T_iQ_i}$ expressed as a credible interval.

4.4 Discussion

In this study, we proposed a method that can be used to adjust for the bias in the diet-disease association due to measurement error in reported dietary intake. Besides adjusting for bias, the method can also adjust for confounding and measurement error in the confounder simultaneously. The strength of this method is that an investigator does not necessarily have to conduct a validation study, provided there is valid knowledge on the extent of measurement error in the self-report instruments that are used. Validation studies are usually very costly to conduct. We demonstrated how to combine external validation data with the observed data to adjust for the bias in the association. The method permits an investigator to either use prior information on the correlation between the errors in the dietary intake and the confounder measurements or on the correlation between their true intakes to estimate the covariance between true intakes. In the EPIC study example, the logHR estimate for FV intake adjusted for the bias with the multivariate method differed slightly from the estimate adjusted for the bias with the univariate method. The logHR estimates for cigarette smoking obtained with both bias- adjustment methods were almost the same. The similarity in the performance of the two methods in our example is due to weak negative error correlation assumed in this study, leading to minimal contamination effect of confounder measurement error. Sensitivity analysis, however, shows that the outcome of the two methods differs strongly when one assumes a strong error correlation. Further found through sensitivity analysis is that depending on the assumed magnitude of measurement error in cigarette smoking, the logHR estimate for FV intake can either be greater or smaller than the unadjusted estimate (Day et al., 2004; Marshall et al., 1999; Wong et al., 1999). Notably, the error in cigarette smoking importantly affected the logHR estimate for FV intake, but not vice versa. This could be due to the stronger effect of cigarette smoking than FV intake on mortality and to the lesser measurement error assumed for cigarette smoking. In our method, we assumed there was no proportional scaling bias, as information on the magnitude of this bias was not available for FV intake and number of cigarettes smoked in a lifetime at the time of this study. However, the proposed method can be easily extended to incorporate such information. In most cases there is no exact external information on the validity of self-report instruments. In such cases, the method allows the user to conduct a sensitivity analysis with a range of plausible estimates to explore the extent to which conclusions derived from the study could be influenced by measurement error. The method also allows pin-pointing assumptions that are crucial for drawing the right conclusion, so that future efforts can be directed towards obtaining valid information.

This method, however, has a few limitations. First, we assumed an additive error structure for the DQ. Generally, however, some intake variables might exhibit multiplicative error structure, where the magnitude of measurement error increases with the quantity of intake (Carroll et al., 2006; Guolo and Brazzale, 2008). In a multiplicative error framework, a remedy could be transform the multiplicative error structure to an additive structure and then proceed with the proposed method. Second, the literature-reported data on validity coefficients for FV intake were based not on gold standards but on concentration markers and recall measurements that do not provide direct measures of true intake (Andersen et al., 2005; Slater et al., 2010). Similarly, cotinine used as a marker for cigarette smoking suffers from same limitation (Pickett et al., 2005; Stram et al., 2002). Thus, the validity coefficients for these variables cannot be determined exactly (Natarajan et al., 2010; Stram et al., 2002). Nevertheless, the Bayesian MCMC sampling-based approach used by the proposed method can still account for the uncertainties in the validity coefficients reported from the literature.

With our example, we illustrate two important features of exposure measurement error. First, measurement error in the confounder can cause bias in the diet-disease association even if dietary intake is measured exactly. Second, when several exposure variables are measured with correlated errors, it can be difficult to predict the direction and magnitude of the association between an exposure and outcome of interest.

4.5 Conclusions

In conclusion, the proposed method can be used to adjust for the bias in the diet-disease association provided there is valid prior information on the magnitude of measurement error in the self-report instrument. The method allows the researcher to venture beyond general statements that measurement error in the confounders might have biased the results, because it allows an assessment of the sensitivity of the estimates to different assumptions regarding the structure of the measurement error. Our example illustrates the well-known fact that measurement error in a major risk factor (e.g., smoking) can affect the association estimate of a suspected risk factor (e.g., FV intake).

List of abbreviations

DQ: Dietary Questionnaire; EPIC: European Prospective Investigation into Cancer and Nutrition Study; FV: Fruits and Vegetables; LogHR: Logarithm of Hazard Ratio; MCMC: Markov Chain Monte Carlo.

Appendices

Appendix 1: How to generate validity coefficients from the range of plausible values obtained from the literature data

Using Fisher z-transformation formula, the validity coefficient ($\rho_{T_iQ_i}$) for the ith study variable is transformed as

$$z_i = \frac{1}{2}\ln\left(\frac{1+\rho_{T_iQ_i}}{1-\rho_{T_iQ_i}}\right),\ i = 1, 2 \text{ (1=FV intake, 2=cigarette smoking)} \tag{4.7}$$

where z_i is approximately normally distributed. We denote the lower and upper limits of the reported validity coefficients by r_l and r_u, respectively. We then use the z_i formula (in expression (4.7)) to obtain the corresponding Fisher z-transformed values for the upper and lower limits of the validity coefficient as z_l and z_u, respectively. Further, using the confidence interval formula for a standard normal random variable, we compute the mean μ_{z_i} and the standard deviation σ_{z_i} of z_i as

$\mu_{z_i} = 0.5(z_u + z_l)$ and $\sigma_{z_i} = \frac{1}{2}\frac{(z_u - z_l)}{Z_{\alpha/2}}$, respectively,

where $Z_{\alpha/2}$ is the $(1-\frac{\alpha}{2})$% quantile of a standard normal random variable. With this parameterization, the z_i are generated as $z_i \sim N(\mu_{z_i}, \sigma_{z_i})$.

Subsequently, the generated z_i are transformed back to the validity coefficient using the inverse of Fisher z- transform as

$$\rho_{T_iQ_i} = \frac{\exp(2z_i) - 1}{\exp(2z_i) + 1} \tag{4.8}$$

Appendix 2: Distribution of correlation coefficients. Kernel densities and histograms for the distribution of validity coefficients for fruit and vegetable (FV) intake ($\rho_{Q_1T_1}$) and number of cigarettes smoked ($\rho_{Q_2T_2}$) as reported in the dietary questionnaires, generated from external validation data by assuming the reported lower and upper limits as 0.05 and 0.95 quantiles of the uncertainty distribution, respectively; the distribution of error correlation (ρ_ϵ) was obtained based on the correlation in the observed data and prior information on the plausible sign of ρ_ϵ for FV intake and cigarette smoking as explained in the main text. Note, with the assumed quantile interval, it is possible to get a small positive values for the error correlation as shown in the distribution of ρ_ϵ.

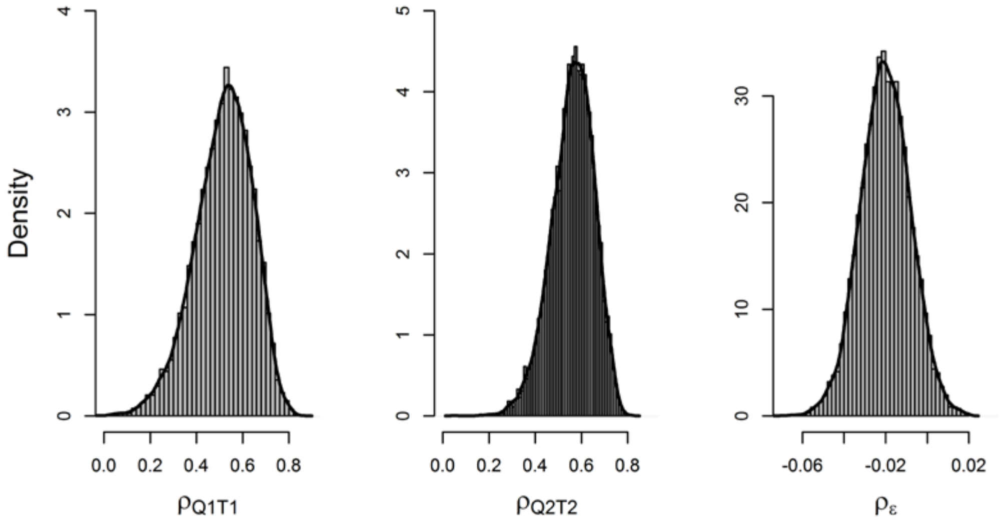

Figure 4.3: Distributions of validity coefficients for FV intake and cigarette smoking as reported on the DQ, and correlation coefficient between the errors.

5 Quantification of Measurement Errors in Accelerometer Physical Activity Data [1]

Abstract

Regular physical activity (PA) is associated with good health. It is, however, difficult to measure PA exactly with an accelerometer in free-living individuals. Measurement errors can bias the associations between PA and health outcomes and lead to loss of statistical power to detect associations. We validated the GT3X accelerometer against doubly labelled water in a study of Dutch adults and compared two prediction equations used to summarize accelerometer data. We estimated: bias in the mean level of PA, correlation coefficient between measured and true PA values (validity coefficient, which quantifies loss of power to detect associations) and attenuation factor (which quantifies bias in the associations), with and without conditioning on age, sex and BMI. We used activity energy expenditure (AEE), total energy expenditure (TEE) and physical activity level (PAL) metrics for PA. The accelerometer underestimated AEE, TEE and PAL by about 450kcal/day, 500kcal/day and 0.3, respectively. Validity coefficients for AEE, TEE and PAL conditional on age, sex and BMI ranged from 0.36 (PAL) to 0.8 (TEE) and for attenuation factors from 0.5 (PAL) to 0.8 (TEE). Results from the two prediction equations were similar. Accelerometer measures perform better for determining TEE than for determining AEE and PAL.

[1] Based on: Agogo, G.O., van der Voet, H., Hulshof, P.J.M., van 't Veer, P.,et al. Quantification of Measurement Error in Accelerometer activity Data (submitted).

5.1 Introduction

Regular physical activity (PA) is associated with good health (Plasqui and Westerterp, 2007; WHO, 2015). PA involves body movement produced by skeletal muscles resulting in energy expenditure above resting levels (Hills et al., 2014; Plasqui, Bonomi and Westerterp, 2013). The health benefits associated with PA are assessed by considering an individual's average physical activity level over a long period of time (hereafter, usual activity) (Nusser et al., 2012). Ideally, usual activity would be measured without error in free-living individuals. PA, however, is difficult to measure exactly under free-living conditions. Thus, usual activity measurements from the accelerometer are subject to error.

The three main effects of measurement error in usual activity are: bias in the mean level of PA, loss of statistical power to detect association between PA and a health outcome, such as obesity (Carroll et al., 2006), and bias in the PA-health outcome association. These three effects of measurement error can be quantified with the mean discrepancy between true and measured activity level in the study population, with the correlation coefficient between measured and true activity level (hereafter, validity coefficient), and with the attenuation factor, respectively (Ferrari et al., 2007; Tooze et al., 2013).

Physical activity contributes to total energy expenditure (TEE). The doubly labelled water (DLW) technique is regarded as the gold standard for measuring TEE in a free-living context (Hallal et al., 2013; Hills et al., 2014; Plasqui and Westerterp, 2007). Total energy expenditure is composed of: energy expended at rest, often referred to as basal energy expenditure (BEE); energy expended above resting level due to PA, referred to as activity energy expenditure (AEE), and thermic effect of food (TEF). The DLW technique requires the use of stable water isotopes and use of sophisticated laboratory equipment for estimating isotope enrichments over time in biological samples (blood and urine). The logistics and combined cost of dosing, sampling and analysis limit the use of the DLW technique in large epidemiologic studies (Hills et al., 2014). Consequently, use of more affordable and objective methods for assessing PA is becoming popular. A commonly used technique to assess PA objectively is accelerometry (Hills et al., 2014; Plasqui et al., 2013). It is, however, widely recognized that accelerometers underestimate some physical activities, such as swimming, cycling, sedentary activities and static exercise in free-living individuals (Hills et al., 2014; Leenders et al., 2001; Lyden et al., 2011; Plasqui and Westerterp, 2007; Van Remoortel et al., 2012). Many validation studies on PA, therefore, use the DLW technique to validate the accelerometer for assessing usual activity in free-living individuals (Hills et al., 2014; Plasqui and Westerterp, 2007). The DuPLO is one such a validation study, where PA was assessed with triaxial GT3X accelerometer (Actigraph, Pensacola, Florida) to monitor body acceleration in a triaxial plane. Doubly labelled water was used as a gold standard for assessing TEE in the DuPLO study.

Presently, there are very few studies on the validity of the GT3X accelerometer model. Moreover, studies on other triaxial accelerometer models usually stop at computing the discrepancy between accelerometer and DLW-derived measurements and correlation coefficient between them, for instance, see Van Remoortel et al. (2012). For adequate validity assessment, however, a researcher needs to know the magnitude of the validity coefficient and attenuation factor associated with the use of the accelerometer (Ferrari et al., 2007; Tooze et al., 2013).

We assessed the validity of the accelerometer used in the DuPLO validation study as follows. First, we applied a plausible model to describe the measurement error structure. Second, we estimated the bias in the population mean level of PA, the validity coefficient and the attenuation factor for three metrics used to describe accelerometer-derived activity measurements: AEE, TEE and physical activity level (PAL). Physical activity level, defined as the ratio of TEE to BEE, is commonly used to measure PA and provides an index of relative excess energy output due to PA (Hills et al., 2014). Third, we estimated these measures conditional on some subject characteristics, as this is the type of validity measurement that is relevant in real world epidemiological studies. Fourth, we proposed a calibration method for the accelerometer to reduce the bias in the population mean level of PA. Lastly, we assessed the performance of two prediction equations commonly used to predict AEE from the accelerometer activity data.

5.2 Materials and methods

5.2.1 DuPLO Study

The DuPLO study participants consisted of a sub-sample from the NQplus study–a longitudinal study on diet and health (https://www.wageningenur.nl/en/project/nqplus.htm). The DuPLO study participants were recruited via email invitation, and were all Dutch, aged 20-70 years and living in Wageningen, Ede, Renkum and Arnhem (Trijsburg et al., 2015). The study was approved by the medical ethics committee of Wageningen University. The purpose of the study was explained to the participants and a written informed consent was obtained from each participant. Among the eligible participants, 200 agreed to participate in the DuPLO study (92 men, 108 women). Data were collected from 2011 to 2013.

5.2.2 Assessment of Energy Expenditure with the DLW

Doubly labelled water was used to measure TEE using the two-point protocol (IAEA, 2009). Subjects were not eligible to join the DLW study if they were planning to travel abroad, on energy restricted diet, using diuretics, lactating, pregnant or planning to be pregnant during the study period, and if they were suffering from congestive heart failure, kidney failure or malabsorption. In total, 70 DuPLO participants joined the DLW study. The TEE assessment covered an eleven-day period. A day before the DLW dose, participants were instructed to follow a normal dietary pattern, refrain from alcohol, heavy exercise and exposure to high temperatures, and to stay in a fasting state the evening prior to DLW dosing. At the first visit, weight and height were measured and baseline urine and saliva samples were collected followed by ingestion of a dose of DLW. Subjects received a mixture of 1.8 g 10% enriched $H_2{}^{18}O$ (Centre for Molecular Research Ltd, Moscow, Russia) and 0.12 g 99.8% enriched ^2H_2O (Cambridge Isotope Laboratories, Inc, Andover, MA, USA) per kg body water. Body weights of male and female were assumed to contain 55% and 50% body water, respectively (Chumlea et al., 2002). Additional urine and saliva samples were collected three and four hours post dose. Participants revisited the study centre eleven days after dosing. At the second visit, body weight was re-measured and two samples of urine and saliva were collected with one hour interval between samples. To quantify within-individual variability in DLW measurements, 30 participants were invited for a second visit and came back for samples collection (mean time in between two measurements ~ 5 months). The samples were analysed at the Centre for Isotope Research, Groningen, The Netherlands (Guidotti et al., 2013). Rate of carbon dioxide production (rCO_2) was calculated as follows: rCO_2 (L/day) = (TBW /2.078)(1.01 kO – 1.04 kD) – 0.0246rGf , where TBW is total body water, kO and kD are isotope elimination rates of oxygen and deuterium, respectively, and rGf = 1.05TBW(kO - kD) (Schoeller, Leitch and Brown, 1986). Total energy expenditure from the DLW was calculated using the modified Weir equation: TEE (kcal/day) = rCO_2 (L/day) x (1.1+3.90/RQ), where RQ was assumed to be 0.85 (Weir, 1949). Activity energy expenditure from the DLW was estimated as $\mathrm{AEE} = 0.9 \times \mathrm{TEE} - \mathrm{BEE}$ (Hills et al., 2014), where BEE was predicted from participant's age, sex, height and weight using Henry's equation (Henry, 2005) and thermic effect of food was taken as 10% of TEE (Hills et al., 2014; Neuhouser et al., 2013).

5.2.3 Assessment of Physical Activity with the Accelerometer

A total of 153 individuals (including the DLW study participants) agreed to participate in the accelerometer study. A GT3X (Actigraph, Pensacola, Florida) accelerometer was used to monitor PA in a triaxial plane. At the first DLW visit, the accelerometer together with instructions was given to each of the 70 participants who agreed to join the accelerometer study. Each participant wore the accelerometer for seven days and kept a record of daily activities. Additionally, 83 more individuals who did not participate in the DLW study agreed to wear the accelerometer. Daily AEE was derived from raw accelerometer activity data using two prediction equations: (i) Freedson VM3 combination (2011) that uses data from all the three axes (Sasaki, John and Freedson, 2011), and (ii) Freedson combination (1998) that uses data from one axis only (Freedson, Melanson and Sirard, 1998). Total energy expenditure from the accelerometer was estimated as $\text{TEE} = (\text{BEE} + \text{AEE})/0.9$ (Neuhouser et al., 2013), where BEE was predicted as explained above. In the analysis, we excluded DLW and accelerometer activity data for one participant who had implausibly low TEE values as compared to BEE.

5.2.4 Measurement Error model for Physical Activity

We denote an activity measurement (expressed either as AEE, TEE or PAL) from the DLW for individual i on day j by R_{ij}, the corresponding activity measurement from the accelerometer by A_{ij} and a latent true usual activity for individual i by T_i. We relate A_{ij} and R_{ij} with T_i using a bivariate linear measurement error model as

$$\begin{cases} A_{ij} = \beta_0 + \beta_A T_i + r_{A_i} + \varepsilon_{A_{ij}} \\ \\ R_{ij} = T_i + \varepsilon_{R_{ij}} \end{cases} \text{,where} \quad \begin{cases} r_{A_i} \sim \text{N}\,(0, \sigma^2_{r_A}) \\ \varepsilon_{A_{ij}} \sim \text{N}\,(0, \sigma^2_{\varepsilon_A}), \\ \varepsilon_{R_{ij}} \sim \text{N}\,(0, \sigma^2_{\varepsilon_R}) \end{cases} \tag{5.1}$$

the intercept term β_0 represents overall bias in the accelerometer that is independent of T_i, and is referred to as constant bias; the slope β_A represents average population bias that is related with T_i, and is referred to as proportional scaling bias; β_0 and β_A are jointly referred to as systematic bias terms (Kipnis et al., 2001); r_{A_i} denotes random deviation of an individual's average bias relative to the average bias in the population, and is referred to as person-specific bias (Ferrari et al., 2007; Kipnis et al., 2003), $\varepsilon_{A_{ij}}$ denotes within-individual random deviation from an individual's average bias; $\varepsilon_{R_{ij}}$ represents within-individual random deviation of DLW measurements from true level of usual activity. We further assume independence between random terms in the model, between each random error component and true usual activity, and between replicate measurements from the same instrument. As a result, A_{ij} is distributed as $A_{ij} \sim \text{N}(\beta_0 + \beta_A \mu_T,\ \beta_A^2 \sigma_T^2 + \sigma^2_{r_A} + \sigma^2_{\varepsilon_A})$ with a mean that is biased for true usual activity mean μ_T. True usual activity is distributed as $T_i \sim \text{N}(\mu_T, \sigma_T^2)$. In contrast, R_{ij} is assumed to be unbiased, i.e., the constant bias term is zero, proportional scaling bias term is one and person-specific bias term is zero. Thus, measurement error in R_{ij} is assumed to be purely due to within-person random variation. In epidemiological studies, analyses on relations with PA are mostly done adjusting for individual characteristics such as age, sex and body mass index. In such analysis, the relevant validity measures are those depending on these characteristics. In order to calculate such conditional validity measures, we reparametrize the distribution of T_i as

$$T_i \sim \mathrm{N}\left(\alpha_0 + \alpha_Z^{\mathrm{T}}\mathbf{Z}, \sigma_T^2\right), \tag{5.2}$$

where $\alpha_0 + \alpha_Z^{\mathrm{T}}\mathbf{Z} = \mu_T$, $\mathbf{Z}$ is a vector of covariates consisting of individual characteristic variables with fixed effect parameters α_Z^{T}.The individual characteristic variables considered in this study include BMI, sex and age.

5.2.5 Quantification of Measurement Error

Measurement error can be quantified in terms of the discrepancy between true and measured mean activity, i.e., with the bias. We explored the bias in mean activity measurements from the accelerometer as follows. First, for each subject with two replicate measurements from the accelerometer and the DLW, we plotted the mean activity estimate from the accelerometer versus the unbiased mean activity estimate from the DLW (hereafter, mean plot). Second, for each subject, we plotted the difference between mean activity estimates from both instruments (as a measure of bias) versus the unbiased mean estimate from the DLW in a Bland-Altman plot (Bland and Altman, 1986; Krouwer, 2008; Lim et al., 2015). In the Bland-Altman plot, we computed 95% limits of agreement between the accelerometer and DLW. The 95% limits of agreement, defined as mean difference ± 1.96 standard deviation of the difference, quantify the level of agreement between activity measurements from both instruments. We also explored the structure of measurement error in each instrument separately using Bland-Altman plots. Using parameters in model
(5.1), the mean bias can be estimated as

$$\widehat{\mathrm{bias}} = \hat{\beta}_0 + \left(\hat{\beta}_A - 1\right)\hat{\mu}_T. \tag{5.3}$$

When the bias is substantial, it is useful to calibrate the accelerometer. We proposed the following method to calibrate the accelerometer activity data. We calibrated AEE derived from the accelerometer ($\mathrm{AEE}_{\mathrm{accel}}$, biased) using TEE from the DLW ($\mathrm{TEE}_{\mathrm{dlw}}$, unbiased) and obtain a calibration factor as

$$\alpha = \frac{(0.9\times\mathrm{TEE}_{\mathrm{dlw}} - \mathrm{BEE})}{\mathrm{AEE}_{\mathrm{accel}}},$$

where α is a calibration factor. After calibrating $\text{AEE}_{\text{accel}}$ by α, TEE and PAL were recalculated.

Loss of statistical power to detect a significant association between PA and health outcome due to measurement error in PA can be quantified with validity coefficient (Ferrari et al., 2007). The validity coefficient is the correlation between measured and true level of PA, and can be expressed in terms of measurement error model parameters as

$$\rho_{AT} = \frac{\text{cov}(A,T)}{\sqrt{\text{var}(T)\text{var}(A)}} = \frac{\beta_A\sigma_T}{\sqrt{\beta_A^2\sigma_T^2+\sigma_{r_A}^2+\sigma_{\varepsilon_A}^2}}, \tag{5.4}$$

where ρ_{AT} is usually between zero and one; a value close to zero signifies substantial loss in statistical power. The association between PA and a health outcome might be biased, typically toward the null when PA is measured with error. The bias toward the null is referred to as attenuation (Kipnis et al., 1997). The extent of attenuation can be quantified with the attenuation factor, λ_A. When the relation between measured and true exposure is linear, λ_A is the regression slope of true on measured exposure, and is expressed in terms of model parameters as

$$\lambda_A = \frac{\text{cov}(A,T)}{\text{var}(A)} = \frac{\beta_A\sigma_T^2}{\beta_A^2\sigma_T^2+\sigma_{r_A}^2+\sigma_{\varepsilon_A}^2}, \tag{5.5}$$

where a λ_A value close to zero indicates severe attenuation. To adjust for the bias in a linear health-outcome model, the unadjusted association estimate is divided by λ_A.

5.2.6 Descriptive statistical analyses and model fitting

We summarized mean activity data from the DuPLO study with a mixed model approach due to imbalance in the study design. We estimated error distributions by computing within-subject differences for activity data derived from DLW and accelerometer separately. We explored error distributions with histograms, density plots, Bland-Altman plots, and formally with the Shapiro-Wilk test. The bias in the accelerometer was explored with mean plot and Bland-Altman plot, as previously described. We subsequently fit the measurement error model using a maximum likelihood method with Newton-Raphson optimization technique and adaptive

Gaussian quadrature with 10 quadrature points. The method was implemented in `SAS` version 9.3 using `NLMIXED` procedure.

5.3 Results

Table 5.1 presents summary measures for relevant study variables in the DuPLO study. On average, male participants were older by 4.6 years, heavier by 13.2 kg, taller by 0.11 m, and with larger mean body mass index by 1.1 kg/m^2 than their female counterparts.

Table 5.1: Overall and Sex-specific mean (standard deviation) and number of observations (N) for subject characteristics in the DuPLO study, Netherlands, 2011-2013

Variables	Overall (N=200)	Male (N=92)	Female (N=108)
	mean (SD)	mean (SD)	mean (SD)
Age, years	55.7 (10.5)	58. 2(9.3)	53.6 (11.0)
Weight, kg	76.0 (14.2)	83.1 (12.9)	69.9 (12.3)
Height, m	1.73 (0.08)	1.79 (0.06)	1.68 (0.06)
BMI, kg/m^2	25.2 (0.04)	25.8 (3.58)	24.7 (4.06)

In Table 5.2, the mean and standard deviation for different energy expenditure metrics are presented. Regardless of the prediction equation used for the accelerometer activity data, AEE, TEE and PAL values derived from the DLW are greater on average and with larger variability (large standard deviation) than their counterparts from the accelerometer. For instance, the accelerometer underestimates mean AEE by about 50% as compared with the DLW. Comparing results from the two prediction equations, AEE predicted with Freedson VM3 Combination (2011) are larger on average and more variable than those predicted with Freedson Combination (1998) equation; this is expected since the latter ignores activity data recorded on two of the three axes of the accelerometer.

Table 5.2: Mean (standard deviation) of BEE (in kcal/day), AEE (in kcal/day), TEE (in kcal/day) and PAL in the DuPLO study, Netherlands, 2011-2013

	Overall	Male	Female
	mean (SD)	mean (SD)	mean (SD)
BEE§, kcal/day	1508.4 (245.9)	1700.1 (200.5)	1345.1 (140.5)
DLW AEE	882.5 (281.2)	1050.5 (264.7)	739.3 (295.2)
ACC AEE, Freedson VM3 ('11) [a]	453.9 (124.1)	486.0 (115.5)	426.6 (131.4)
ACC AEE, Freedson ('98) [b]	414.6 (114.7)	455.5 (106.7)	379.8 (121.5)
DLW TEE	2678.6 (343.2)	3047.6 (323.1)	2364.3 (360.3)
ACC TEE [a]	2185.1 (192.8)	2423.0 (179.4)	1982.5 (204.2)
ACC TEE [b]	2141.7 (186.9)	2389.5 (173.9)	1930.7 (197.9)
DLW PAL	1.75 (0.20)	1.80(0.19)	1.71 (0.21)
ACC PAL [a]	1.45 (0.09)	1.43 (0.08)	1.46 (0.09)
ACC PAL [b]	1.42 (0.08)	1.41 (0.07)	1.42 (0.08)

Abbreviation: BEE , basal energy expenditure, DLW, doubly labelled water; ACC, accelerometer; AEE, activity energy expenditure; TEE, total energy expenditure; PAL, physical activity level expressed as a ratio of TEE to BEE,

[§] BEE predicted from age, sex and weight using Henry's equation,

[a] Accelerometer-derived AEE, TEE and PAL, where AEE is predicted with Freedson VM3 (2011) combination equation;

[b] Accelerometer-derived AEE ,TEE and PAL, where AEE is predicted with Freedson Combination (1998) equation.

Figure 5.1 displays Bland-Altman plots for AEE derived from the accelerometer (a), DLW (b) and from both instruments (d); also shown is the scatterplot of the mean AEE estimate from the accelerometer versus the mean estimate from the DLW. In Figure 5.1 (a) and (b), the scatter plots appear to be spread randomly and do not show any discernible trend. Lack of trend in the scatter plots suggests that the magnitude of errors in the accelerometer and DLW do not depend on the mean level of AEE, i.e., the errors are additive. Symmetrically distributed within-subject differences explored with histogram and density plots suggest normally distributed errors (graphs not shown); Shapiro-Wilk test also resulted in statistically non-significant p-values. Figure 5.1 (c) suggests that the AEE for subjects with large mean DLW values are underestimated more with the accelerometer measurements than for subjects with small mean DLW values. The flattened regression slope in Figure 5.1 (c) suggests existence of proportional scaling bias in the accelerometer activity data. The accelerometer underestimates mean AEE by about -441.6 kilo calories per day (Figure 5.1 (d), dotted middle line).

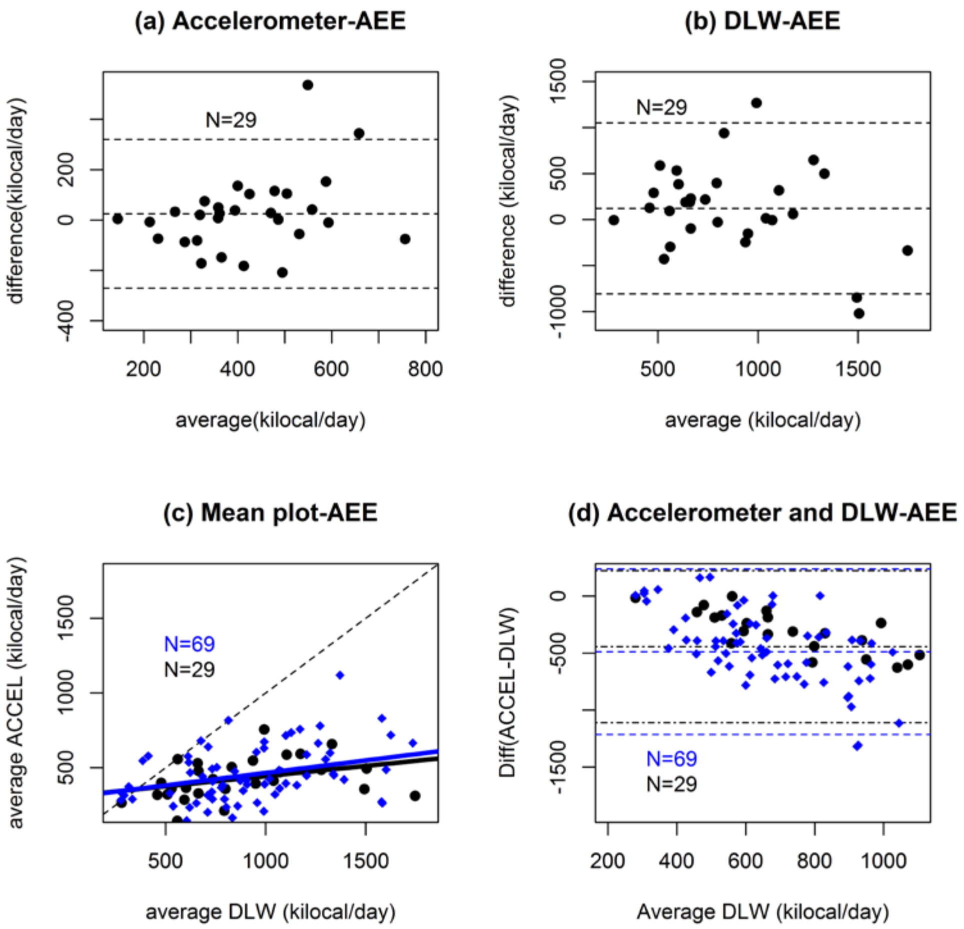

Figure 5.1: Bland-Altman plot for activity energy expenditure (AEE) measurements derived from accelerometer (a) and DLW (b),where within-subject differences are plotted against subject averages, also shown is the mean difference (middle dotted line) and 95% limits of agreement (extreme dotted lines); In (c), subject average AEE measurements from accelerometer are plotted against subject averages from the DLW; In (d) within-subject between-instrument differences in the mean AEE estimate are plotted against the corresponding subject averages from the DLW; blue dots (N=69) refers to first replicate measurements, DuPLO study, Wageningen, Netherlands, 2011-2013.

Based on these findings, we assumed normality and additivity for distributions of within-person errors in the accelerometer and DLW measurements, and systematic bias in the accelerometer measurements. We observed similar findings for TEE and PAL measurements. We consequently fitted the proposed measurement error model unconditionally and by letting true usual activity depend on the subject's age, sex and BMI. Age and BMI were standardized to improve convergence of the model.

Table 5.3 presents parameter estimates of measurement error in the accelerometer. There is evidence of overall bias in the accelerometer that is independent of an individual's level of activity (β_0). For instance, in the model where T_i is predicted conditional on the covariates and the accelerometer–derived AEE is predicted by Freedson VM3 (2011) equation, the mean AEE will be underestimated by 157 kilocalories per day. The estimate of β_A, the bias that depends on an individual's level of activity, is less than one for AEE, TEE and PAL in all cases; this means that usual activity for very physically active individual is underestimated more than less physically active individuals with the accelerometer. There is person-specific bias σ_{r_A} and within-person random error σ_{ε_A} in the accelerometer measurements. From this, it is evident that the accelerometer underestimates mean usual activity. For instance, the accelerometer underestimates mean AEE by about 445 kilo calories per day, when a model conditional on the covariates is used as well as Freedson VM3 (2011) prediction equation (Table 5.4).

Table 5.3: Estimates for parameters of measurement error in the accelerometer and variance of true activity expressed as AEE, TEE and PAL, when true activity is predicted with and without conditioning on subject's age, sex and BMI, where AEE from accelerometer is predicted with Freedson VM3 (2011) combination equation [a] and Freedson Combination (1998) equation [b], DuPLO study, Wageningen, Netherlands, 2011-2013

Metric	True activity level conditional on the covariates†	Intercept of regression of accelerometer activity on true activity β_0 (SE)	Regression slope of accelerometer activity on true activity β_A (SE)	Standard deviation of true activity level σ_T(SE) × 10^3	Standard deviation of person-specific bias in accelerometer σ_{r_A}(SE) × 10^3	Standard deviation of within-person error in accelerometer σ_{ε_A}(SE) × 10^3	Standard deviation of within-person random error in DLW σ_{ε_R}(SE) × 10^3
AEE	Yes [a]	0.157 (0.094)	0.334(0.101)	0.240 (0.048)	0.127 (0.018)	0.103 (0.009)	0.322 (0.033)
	Yes [b]	0.104 (0.092)	0.351(0.098)	0.226 (0.049)	0.110 (0.018)	0.100 (0.009)	0.330 (0.034)
	No [a]	0.103 (0.146)	0.391(0.158)	0.264 (0.058)	0.117 (0.026)	0.103 (0.009)	0.333 (0.040)
	No [b]	0. 093 (0.131)	0.359(0.141)	0.268 (0.058)	0.108(0.025)	0.100 (0.009)	0.332 (0.040)
TEE	Yes [a]	0.149 (0.240)	0.758 0.087)	0.228 (0.033)	0.088 (0.040)	0.109 (0.009)	0.405 (0.033)
	Yes [b]	0.053(0.246)	0.778(0.089)	0.212 (0.032)	0.073 (0.046)	0.107 (0.009)	0.412 (0.033)
	No [a]	0.387 (0.287)	0.670(0.105)	0.480 (0.058)	0.151 (0.053)	0.110 (0.009)	0.371 (0.045)
	No [b]	0.359 (0.279)	0.664(0.102)	0.482 (0.058)	0.155 (0.051)	0.108 (0.009)	0.368 (0.044)
PAL	Yes [a]	0.827 (0.317)	0.352(0.178)	0.164 (0.035)	0.089 (0.018)	0.076 (0.006)	0.226 (0.025)
	Yes [b]	0.971 (0.239)	0.254(0.134)	0.164 (0.035)	0.083 (0.013)	0.072 (0.006)	0.224 (0.023)
	No [a]	0.669 (0.374)	0.441(0.211)	0.161 (0.042)	0.080 (0.020)	0.076 (0.006)	0.234 (0.027)
	No [b]	0.768 (0.323)	0.368(0.182)	0.162 (0.042)	0.073 (0.018)	0.072 (0.006)	0.234 (0.027)

† Whether a subject's true activity is predicted by conditioning on age, sex and BMI or not; AEE, activity energy expenditure; TEE, total energy expenditure; PAL, physical activity level (PAL=TEE/BEE);

[a] AEE is predicted with Freedson VM3 (2011) combination equation that uses activity data from the three axes;

[b] AEE is predicted with Freedson Combination (1998) equation that uses activity data from one axis only.

Based on the validity coefficient estimates shown in Table 5.4, there is a substantial loss of statistical power in testing the association between PA and any health outcome, when PA is expressed either as AEE or PAL and measured by accelerometer. For instance, to attain the required power to detect a significant association when ρ_{AT} is 0.44, the sample size of accelerometer study should be about five times as large as (i.e.,$1/\rho_{AT}^2 = 1/0.44^2$) the sample size that would be required if AEE were measured exactly. There is, however, a small loss in statistical power when PA is expressed as TEE, i.e., the sample size should be 1.6 times as large.

Table 5.4: Estimates for validity coefficient, attenuation factor and mean bias for accelerometer –derived activity measurements, when true activity is predicted with and without conditioning on subject's age, sex and BMI, where AEE from accelerometer is predicted with Freedson VM3 (2011) combination equation [a] and Freedson Combination (1998) equation [b], DuPLO study, Wageningen, Netherlands, 2011-2013

Metric	True activity level conditional on the covariates †	Correlation between accelerometer and true activity ρ_{AT} (SE)	Attenuation factor λ_A (SE)	Mean bias in accelerometer measurements in kcal/day ×10^3
AEE	Yes [a]	0.44 (0.127)	0.59 (0.229)	-0.445 (0.091)
	Yes [b]	0.47 (0.120)	0.63 (0.240)	-0.484 (0.095)
	No [a]	0.55 (0.154)	0.78 (0.232)	-0.458 (0.043)
	No [b]	0.55 (0.151)	0.83 (0.253)	-0.499 (0.043)
TEE	Yes [a]	0.78 (0.089)	0.80 (0.177)	-0.502 (0.096)
	Yes [b]	0.79 (0.094)	0.80 (0.186)	-0.545 (0.088)
	No [a]	0.86 (0.068)	1.12 (0.133)	-0.513 (0.047)
	No [b]	0.86 (0.067)	1.12 (0.133)	-0.558 (0.047)
PAL	Yes [a]	0.44 (0.194)	0.55 (0.263)	-0.321 (0.029)
	Yes [b]	0.36 (0.187)	0.50 (0.302)	-0.349 (0.039)
	No [a]	0.54 (0.171)	0.66 (0.217)	-0.323 (0.029)
	No [b]	0.50 (0.174)	0.68 (0.249)	-0.354 (0.029)

† Whether a subject's true activity is predicted conditional on age, sex and BMI or not; AEE, activity energy expenditure; TEE, total energy expenditure; PAL, physical activity level (PAL=TEE/BEE);

[a] AEE is predicted with Freedson VM3 (2011) combination equation that uses activity data from the three axes; [b] AEE is predicted with Freedson Combination (1998) equation that uses activity data from one axis only

There will be a sizable attenuation in the PA-outcome associations, when PA measurement from the accelerometer is expressed either as AEE or PAL (Table 5.4). For instance, when AEE measurements from the accelerometer, with a λ_A value of 0.59 are used, a true relative risk of 0.6 for beneficial effect of regular PA would be observed as 0.74 ($0.6^{0.59}$); when PAL measurements from the accelerometer, with a λ_A value of 0.55 are used, one would observe a relative risk of 0.76. When PA is expressed as TEE with a λ_A value of 0.8, one would observe a modest attenuation with a relative risk of 0.66. Regardless of the metric used, the observed relative risk would be weaker than the true relative risk. It is clear that when we look at true activity conditional on subject's age, sex and BMI, this affects the magnitude and level of precision of parameter estimates from the measurement error model. Both prediction equations for AEE from the raw accelerometer data result in similar magnitudes for the parameter estimates in the model, validity coefficients and attenuation factors, but with a greater magnitude of bias for Freedson Combination (1998) as expected.

Table 5.5 presents results before and after calibrating the AEE measurements from the accelerometer by a calibration factor (α) of 1.9, estimated with the proposed calibration approach, and recalculating TEE and PAL. As expected, the proposed calibration method reduces bias in the mean usual activity substantially. Specifically, the mean bias is reduced by 93% for AEE, 90% for TEE and by 95% for PAL. The calibration method, however, does not reduce loss of statistical power and attenuation of PA-health outcome associations.

Table 5.5: Parameter estimates for the measurement error model, validity coefficient, and attenuation factor and mean bias (per 1000 kcals/day) estimates AEE, TEE and PAL measurements from the accelerometer using Freedson VM3 (2011) combination equation, with and without calibrating the accelerometer activity data, DuPLO study, Wageningen, Netherlands, 2011-2013

Metric	Accelerometer AEE calibrated[a]	Correlation between accelerometer and true activity ρ_{AT} (SE)	Attenuation factor λ_A (SE)	Mean bias in accelerometer measurements in kcal/day × 10^3
AEE	Not calibrated	0.44 (0.127)	0.59 (0.229)	-0.445 (0.091)
	Calibrated	0.44 (0.127)	0.31 (0.121)	-0.031 (0.050)
TEE	Not calibrated	0.78 (0.089)	0.80 (0.177)	-0.502 (0.096)
	Calibrated	0.63 (0.084)	0.44 (0.111)	-0.050 (0.042)
PAL	Not calibrated	0.44 (0.194)	0.55 (0.263)	-0.321 (0.029)
	Calibrated	0.44 (0.194)	0.29 (0.139)	-0.015 (0.015)

[a] Whether accelerometer-derived activity data are calibrated to reduce mean bias or not; AEE, activity energy expenditure; TEE, total energy expenditure; PAL, physical activity level.

5.4 Discussion

We assessed the validity of a triaxial accelerometer (GT3X) in the DuPLO study, by calculating the bias in the mean activity level, the correlation coefficient between measured and true activity level and the magnitude of attenuation in the association between physical activity and a health outcome of interest. The accelerometer underestimated TEE by about 18% on average as compared with the DLW, which is within the 95% confidence interval reported in a review study by Van Remoortel et al. (2012) and consistent with findings from other similar studies (Plasqui et al., 2013; Plasqui and Westerterp, 2007). Similarly, the accelerometer underestimated AEE by about 49%, consistent with findings from Leenders et al. (2001) for a TriTrac accelerometer that monitors body acceleration in a triaxial plane. Despite the plausibility of our study findings in context of the literature, the magnitude of underestimation of PA in free-living individuals with the accelerometer in our study seemed more severe than in most literature studies such as in Van Remoortel et al. (2012). The observed severe underestimation in both our study and those of others could be due to a number of reasons; for instance, too simplistic prediction equations for accelerometer-derived AEE (Plasqui et al., 2013), and low sensitivity of the accelerometer to monitor sedentary activities, bicycling and static exercise, such as fidgeting, especially when worn at the waist (Hills et al., 2014; Swartz et al., 2000). Notably, the DuPLO study participants bicycled regularly and failure to monitor bicycling could have resulted in more severe underestimation in the DuPLO study than in studies conducted in other countries. Without monitoring fidgeting alone, an individual’s daily TEE could be underestimated by up to 800 kilo calories per day (Leenders et al., 2001).

The DuPLO study analysis revealed that the accelerometer underestimated true mean activity, especially for physically very active individuals. Validity coefficient estimate for AEE and PAL suggested substantial loss of statistical power to detect PA-outcome associations. We further found that the association would be attenuated when accelerometer-derived activity data is used to estimate the association. Previous studies showed similar findings when physical activity was assessed with the questionnaires (Ferrari et al., 2007). Presently, the study by Ferrari et al. (2007) is the closest to our study in terms of design and objectives. It is, however, difficult to compare our results quantitatively with those of Ferrari et al. (2007). The difficulty in comparison is due to two reasons. First, the authors expressed physical activity in log-transformed MET hours per week as opposed to untransformed kilo calories per day in our study. Second, they assumed PA logs as the reference measure as opposed to DLW in our study.

In our analysis, conditioning on the subject’s characteristics influences the validity measures, in line with findings from the literature (Ferrari et al., 2007). Subject’s sex, age and body mass index contributed to between-individual variability in activity level. With a higher between-individual variability in true values, the correlation between true and measured values increases. In epidemiologic analysis, it is common to either adjust for the effects of these individual characteristics or stratify the analysis accordingly. Therefore, validity measures, conditioned on these covariates are more relevant in reality.

The magnitudes of validity coefficient and attenuation factor estimates were similar irrespective of the prediction equation used to predict AEE from the accelerometer data. The similarity in the estimates suggests minimal contribution of activity data recorded on all the three axes over activity data recorded on one axis. This finding is in line with previous studies that showed minimal improvement when AEE was measured with a triaxial accelerometer over a uniaxial one (Hills et al., 2014), or by using one prediction equation over the other (Crouter, Churilla and Bassett, 2006; Leenders et al., 2001).

This study provides an in-depth description of measurement error in the accelerometer activity data and essential components of plausible error structure for the GT3X accelerometer model. The proposed calibration approach is intuitive and corrects for the mean bias in the accelerometer measurements in the DuPLO study population. However, whether this applies to other populations needs further investigation.

This study had a few limitations. First, its external validity is limited because DuPLO participants were of similar ethnicity, living in the same region and were all adults. Thus, generalizing the study findings to other different populations might be misleading. Second, BEE was predicted with an equation, which could result in extra error. There are more reliable but expensive methods to measure BEE, such as indirect calorimetry. Finally, because there is no universally recognized gold standard for measuring AEE, we assumed the respective estimates derived from DLW as the gold standard; the implication of this assumption on the validity measures requires further investigation.

In conclusion, the accelerometer underestimated mean usual activity in the DuPLO study population. Given the measurement error model used in this study, there would be substantial loss in statistical power to detect associations and there would be bias in the association between physical activity and a health outcome, when physical activity level or activity energy expenditure is assessed with the GT3X accelerometer. However, validity is better for measurement of TEE.

List of abbreviations

AEE: Activity Energy Expenditure, BEE: Basal Energy Expenditure, DLW: Doubly Labelled Water, PA: Physical Activity, PAL: Physical Activity Level, TEE: Total Energy Expenditure, TEF: Thermic Effect of Food.

6 General discussion

Measurement errors in exposure variables can bias the associations between exposures and outcomes (Carroll, Freedman and Kipnis, 1998; Kipnis et al., 2003). Several methods have been developed to handle measurement error, but generally require validation studies with multiple replicates of unbiased exposure measurements (Buonaccorsi, 2010; Carroll et al., 2006; Fuller, 2006). In this thesis, we examined the possibility of adapting some of the methods to adjust for measurement error in other study designs without multiple replicate measurements, develop a method to handle measurement error in studies without an internal validation data, and apply measurement error model to a new validation dataset.

Thus, the research in this thesis focused on the following three key aspects:

(i) A method to adjust for the exposure measurement error in the presence of a single-replicate validation study for episodically consumed foods (chapters 2 and 3),

(ii) A method to adjust for correlated measurement errors in multiple exposures when there is no internal validation study (chapter 4),

(iii) Validation of an accelerometer when used to measure physical activity in free-living individuals (chapter 5).

In this chapter, the main findings are summarized in Table 6.1 below. These main findings are discussed in a general context under each of the three key aspects shown above. Next, study limitations and implications are discussed followed by suggestions for improvement and potential areas for future research. The chapter ends with concluding remarks.

Table 6.1: The main findings and main message from this thesis

Main findings	**Main message**
Chapters 2 and 3	
(i) Calibration using inappropriately specified functional forms of continuous covariates in a regression calibration can worsen the bias in the exposure-outcome associations (ii) For data with excess zeroes, the use of two-part regression calibration is theoretically optimal, but the one-part calibration model also appears robust. (iii) Reducing a standard calibration model may improve the performance in adjusting for the bias in the exposure-outcome associations, but the improvement is minimal for large studies (iv) The performance of two-part calibration model was minimally affected by the magnitude of correlation between the probability of a positive response and the conditional response given a positive response value	• When a single-replicate validation study with zero-inflated reference measurements is available, a suitably specified regression calibration can be used to adjust for the bias in the exposure-outcome associations
Chapter 4	
(i) If the confounder is strongly linked with the outcome, measurement error in the confounder can be more influential than measurement error in the exposure in causing bias in the exposure-outcome associations (ii) In a sensitivity analysis, it is shown that in the presence of mismeasured confounders, the exposure-outcome associations can still be biased, even when the exposure is measured without error (iii) In the presence of correlated measurement errors, the exposure-outcome associations can be attenuated, inflated or can even reverse directions	• When there is no internal validation study, carefully extracted external validity data for self-report instruments can be useful in adjusting for the bias in exposure-outcome associations • The proposed method is useful in conducting sensitivity analysis on the effect of confounder measurement error and error correlation on the observed exposure-outcome association
Chapter 5	
(i) When an accelerometer, such as GT3X, is used to monitor physical activity in free-living individuals, it is likely that the mean level of physical activity is underestimated, the associations between physical activity and health outcomes is biased and there is loss of statistical power to detect associations	• When a "reference" instrument used to validate a main-study instrument is itself marred by substantial error, the true effect of measurement error in the main-study instrument will be misrepresented

6.1 A method to adjust for exposure measurement error in the presence of a single-replicate validation study for episodically consumed foods

6.1.1 Main findings

In chapter 2 we adapted the regression calibration method to handle zero-inflation and skewness in the response using a single-replicate validation study. Most measurement error correction methods were originally developed for a multiple-replicate validation study. A method that uses a single-replicate validation data to adjust for measurement error in episodically consumed foods was lacking. Zero inflation and skewness are common distributional characteristics of episodically consumed dietary intakes.

In a simulation study presented in chapter 3, we showed that the magnitude of the correlation between an individual's probability of consumption and the amount consumed on consumption days (cross-part correlation) did not matter much, when the adapted regression calibration method was used to adjust for intake measurement error. In the proposed method, the probability of consumption and the amount consumed on consumption days were assumed to be independent given the covariates in the calibration model. In practice, it is not unlikely that the two consumption components: consumption probability and consumed amount are still correlated, even if corrected for the covariates. This part of correlation is due to random component that is not explained by the covariates. This conditional independence assumption is not required in the existing methods that use multiple replicate measurements (see Table 1.1 in chapter 1).

Furthermore, in chapter 3, it is shown that the adequacy of measurement error correction can be influenced by how the adapted regression model is specified. The estimated exposure-outcome association can change substantially when a calibration model is insufficiently specified, mainly with respect to functional forms of continuous covariates. In the regression calibration literature, relations are usually assumed to be linear. In practice, however, this is not always adequate, especially when dealing with dietary exposures that are often skewed. Failure to model nonlinear terms correctly in the calibration model might lead to outlying predicted values when predictor variables for subjects in the main study lie outside the variable space of the sample used to fit the prediction model. When these extreme predictions are used as calibrated values for the exposure in a model that relates the exposure with an outcome, the estimation of the association can be distorted (Greenland, 1989). The impact of the distortion can be in any direction, such that the estimated association parameter that is adjusted for exposure measurement error may appear smaller than the association estimate that ignores exposure measurement error, even for the univariate case where the exposure variable is measured with random error.

The performance of the calibration model can be brought down by model complexity. The results in chapter 2 suggest that non-significant covariates can lead to over fitting and hence poor predictive power. As a result, there will be increasingly larger part of the variable space in the main study that is not adequately covered by the data used to fit the calibration model. Thus, to recover parsimony, the complexity of the calibration model can be reduced by using a suitable variable reduction method such as backward elimination (see chapters 2 and 3).

6.1.2 Limitations

6.1.2.1 Assumptions in the reference instrument

Assumptions regarding the reference measurements in regression calibration are seldom satisfied in practice, because reference instruments do not exist for most dietary variables (Willet, 1998). To use regression calibration, the following assumptions are made on the reference measurements: (a) that they are an unbiased measure for the true exposure and, (b) that they are measured with errors that are uncorrelated with (i) the true exposure and (ii) the measurement errors in the main-study instrument (Kipnis et al., 2001), and (c) that the measurement errors are nondifferential (Carroll et al., 2006). If error is nondifferential, the measured exposure does not provide extra information about the outcome over what is contained in the true exposure. For episodically consumed foods, it has been shown that a single-day intake from a short-term instrument, even if unbiased, might be a very imprecise representation of true long-term intake, due to sizable between-day variation and excess zeroes (Kipnis et al., 2009; Tooze et al., 2006). Therefore, it is important to assess the adequacy of the adapted calibration method. Such an assessment can be done via a simulation study as shown in chapter 3.

6.1.2.2 Correlated systematic bias in the 24 hour recall and dietary questionnaire

In chapter 2, the 24 hour recall (hereafter, 24HR) was used as the reference instrument. However, when used to measure protein and energy intakes, 24HR has been shown to be marred by systematic bias and with errors that are correlated with the errors in the questionnaire (Kipnis et al., 2003). With the vegetable intakes example in chapter 2, it is common for individuals to over-report their intakes on the 24HR, because vegetable intake is considered good for their health. Furthermore, it is likely for individuals who over-report their vegetable intakes on the 24HR to also over-report their intakes in the dietary questionnaire (hereafter, DQ), leading to positively correlated errors. To understand the implications of the bias in the "reference" instrument in the study conclusions, we use the following measurement error model for the DQ (Q_i) and the 24HR (R_i) as an example:

$$\begin{aligned} Q_i &= \beta_{Q0} + \beta_Q T_i + s_{Q_i} + \varepsilon_{Q_i}, \\ R_i &= \beta_{R0} + \beta_R T_i + s_{R_i} + \varepsilon_{R_i}, \end{aligned} \tag{6.1}$$

where β_{Q0} and β_{R0} quantify overall constant bias for the DQ and 24HR, respectively, β_Q and β_R are slopes that quantify intake-related bias for the DQ and 24HR (Kipnis et al., 2001); s_{Q_i} and s_{R_i} are person-specific bias for the DQ and 24HR that are assumed to be independent of true intake T_i, have means zero, variances σ_q^2 and σ_r^2, respectively, and are correlated with the correlation coefficient ρ_{qr}. The person-specific biases s_{Q_i} and s_{R_i} describe the fact that two individuals who consume the same amount of food will systematically report their intakes differently (Carroll et al., 2006); ε_{Q_i} and ε_{R_i} are within-person random errors for the DQ and 24HR with means zero and variances $\sigma^2_{\varepsilon_Q}$ and $\sigma^2_{\varepsilon_R}$, respectively, that are assumed to be independent of each other and of other terms in the model. In regression calibration, 24HR is assumed as a valid instrument such that $\beta_{R0} = 0,\ \beta_R$ =1 and σ_r^2 = 0. In the case of only a single-replicate measurement, it is impossible to disentangle person-specific bias s_{Q_i} and within-person random error ε_{Q_i} in Q_i; the same applies to R_i. If the 24HR is a valid reference instrument, i.e., no intake-related bias ($\beta_R = 1$) and person-specific bias ($\sigma_r^2 = 0$), then the true attenuation factor (λ_T) that quantifies the bias in the exposure-outcome association is given by

$$\lambda_T = \frac{\mathrm{cov(Q,T)}}{\mathrm{var(Q)}} = \frac{\beta_Q}{\beta_Q^2+\sigma_q^2/\sigma_T^2+\sigma^2_{\varepsilon_Q}/\sigma_T^2}. \tag{6.2}$$

Otherwise, the observed attenuation factor (λ_R) is given by $\lambda_R = \mathrm{cov}(R,Q)/\mathrm{var}(Q) = \beta_Q\beta_R\sigma_T^2 + cov(s_{Q_i}, s_{R_i})/var\ (Q)$. The λ_R can be re-expressed in terms of the true attenuation factor λ_T (Kipnis et al., 2003) as

$$\lambda_R = \lambda_T\left(\beta_R + \frac{1}{\beta_Q}\rho_{qr}\sqrt{\frac{\sigma_q^2}{\sigma_T^2}\frac{\sigma_r^2}{\sigma_T^2}}\right). \tag{6.3}$$

The magnitude of bias in the attenuation factor λ_R, therefore, depends on intake-related bias for the 24HR (β_R) and DQ (β_Q), the correlation between their person-specific biases (ρ_{qr}) and the variances of their person-specific biases relative to the variance of true intake (σ_q^2/σ_T^2 and σ_r^2/σ_T^2). In Table 6.2, we illustrate, with a hypothetical numerical example, the effect of correlation between person-specific biases in the 24HR and DQ, and the ratio of their variances to true variance on the attenuation factor. These parameter values are close to those for energy intake presented in the OPEN study (Kipnis et al., 2003). For illustration, we assume that the values used here are transferable to the case of vegetable intake measurements. With this example, if 24HR is a valid reference instrument ($\beta_R = 1, \sigma_r^2 = 0$), the attenuation factor will be estimated exactly ($\lambda_R = \lambda_T$, first row). However, the attenuation factor will be underestimated when the correlation coefficient between person-specific biases is close to zero (e.g., if $\rho_{qr} = 0.05, \lambda_R = 0.27 < \lambda_T$, second row) or when the variance of the person-specific bias is reduced relative to the true variance (e.g., if $\sigma_r/\sigma_T = 0.3, \lambda_R = 0.25 < \lambda_T$, fourth row). Conversely, the attenuation factor will be overestimated when

ρ_{qr} becomes more positive (e.g., if $\rho_{qr} = 0.15, \lambda_R = 0.45 > \lambda_T$, third row) or when β_R is close to a theoretical value of one for no intake-related bias (e.g., if $\beta_R = 0.8, \lambda_R = 0.51 > \lambda_T$, fifth row).

Table 6.2: The effect of correlation between person-specific biases and the ratio of their variances to true variance on the attenuation factor

λ_R	λ_T	β_R	β_Q	ρ_{qr}	$\sqrt{\sigma_q^2/\sigma_T^2}$	$\sqrt{\sigma_r^2/\sigma_T^2}$
0.30	0.30	1.0	0.4	0.00	2	0.0
0.27	0.30	0.6	0.4	0.05	2	1.2
0.45	0.30	0.6	0.4	0.15	2	1.2
0.25	0.30	0.6	0.4	0.15	2	0.3
0.51	0.30	0.8	0.4	0.15	2	1.2

With this example, if there is intake-related bias in the 24HR and that person-specific bias in the 24HR is correlated with person-specific bias in the DQ, the observed attenuation factor will misrepresent the true attenuation factor. As a result, there will still be bias in the estimated exposure-outcome association that is calibrated with the observed attenuation factor λ_R. Nevertheless, it is likely that the deattenuated association will be closer to the truth than the crude association.

6.1.2.3 Nondifferential error assumption

The assumption of nondifferential error in our case seemed to be plausible. We used a nutritional study, where the outcome of interest (all-cause mortality) occurred (mostly) many years after the assessment of long-term dietary intake. Therefore, it is reasonable to assume that dietary intake reported in the DQ does not provide extra information on all-cause mortality over what is provided by true long-term dietary intake. Nevertheless, a possibility of differential error cannot be ruled out, as measurement error in reported intake could be related to subject characteristics (not included in the modelling) that predict mortality. If measurement errors in the DQ are indeed differential, then the results from our method would be biased.

6.1.3 Suggestions for possible improvement

Besides parametric transformations used to describe nonlinear relations between the response in the calibration model and skewed covariates, other methods can be used such as flexible fractional polynomials approach (Royston and Sauerbrei, 2003, 2004, 2005). The fractional polynomials approach encompasses a wide range of covariate transformations of the form x^k, where x could be DQ intake and k is a set of integer and non-integer values (Royston and Sauerbrei, 2008). Notably, in regression calibration, the relation between the reference measurements and the biased measurements are assumed to be monotonically increasing. Thus, the suitability of higher order fractional polynomials is limited when a monotonically increasing form of relation is required. Nonlinear relations can also be handled with non-parametric smoothing techniques such as loess, kernel and splines (Carroll, Delaigle and Hall, 2009; Hastie and Tibshirani, 1999; Hastie et al., 2009), or semi-parametrically with generalized additive modelling approach (Hastie and Tibshirani, 1999). To achieve the desired form of relation with these non-parametric methods requires some penalty constraints; for instance, with penalized splines (Hastie and Tibshirani, 1999; Hastie et al., 2009).

6.1.4 Our method in relation to other measurement error methods

In this thesis, regression calibration was adapted to handle a validation study design with single-replicate measurements that are characterized by many zeroes. It would be useful to determine whether other existing methods, such as likelihood-based, Bayesian, simulation extrapolation and multiple imputation, could be adapted to handle such study designs. In regression calibration, only the conditional expectation of true exposure is estimated. Thus, a validation study with only a single unbiased measurement and a possibly biased exposure measurement from the main study is required. However, other measurement error correction methods do not only need specification of the mean but also the variance of true exposure. For example, to apply likelihood or Bayesian methods, the distribution for true exposure must be specified (Carroll et al., 2006). This can only be estimated from a validation study with at least two replicates of unbiased measurements. Likewise, to apply simulation extrapolation (SIMEX) method, the measurement error variance must be known requiring multiple replicate measurements. To apply multiple imputation method, the distribution of true exposure given the observed data, including the outcome data, must be specified. This conditional distribution can only be estimated from a validation study with data on the study outcome and enough subjects with multiple replicates of the reference measurements in both subjects with and without the outcome (Keogh and White, 2014). The conclusion therefore is that in single-replicate studies, it is not feasible to apply these other methods.

6.2 A method to adjust for measurement error in multiple exposures measured with correlated errors when there is no internal validation study

6.2.1 Main findings

Measurement error can occur not only in the exposure of interest but also in the confounders. In chapter 4, we showed how to adjust for measurement errors in both exposure and confounder variables when there is no internal validation study. The proposed method combines prior information on the validity of self-report instruments with the observed data to adjust for the bias in the associations. The proposed method can improve quantitative inference in epidemiological studies, where the problem is currently mostly addressed only with a qualitative remark in the discussion section. We illustrated the method with an epidemiological study with single-replicate DQ intake data per subject and used prior information on the validity of the DQ for measuring both the exposure and confounder and prior assumptions on the sign of the correlation between the respective measurement errors. Our example illustrates that (a) a confounder can be a serious problem when (i) measured with substantial errors, (ii) strongly correlated with the exposure and (iii) strongly linked with the outcome, and (b) that measurement error in the confounder can bias the exposure-outcome association, even if the exposure is measured exactly.

6.2.2 Explanation of the findings

When a confounder is measured with error, the observed association may be smaller or larger than the true association, and can even reverse its sign. Though not shown in this thesis, an exposure variable with no association with the outcome may appear to have a sizable effect due to confounding by other variables (Marshall et al., 1999). An intuitive explanation is that when both exposure and confounder are measured with error, they will each adopt a fraction of the other's effect, such that the observed associations will be contaminated. The degree of contamination will depend upon the strength of the correlation between the two (true) variables: exposure and confounder, their variances, and the correlation and variances of their measurement errors. We explain the contamination effect due to confounding with the so-called attenuation-contamination matrix (Freedman et al., 2011; Rosner et al., 1990) given by

$$\Lambda = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{bmatrix}, \tag{6.4}$$

where the diagonal elements (λ_{11} and λ_{22}) are the attenuation factors and the off-diagonal elements (λ_{21} and λ_{12}) are often referred to as the contamination factors (Freedman et al., 2011). The elements of Λ can be expressed in terms of variances and covariances. To do this, we denote true values for the exposure and the confounder by T_1 and T_2, respectively, and assume no systematic bias in their measured values denoted by Q_1 and Q_2, respectively. Using the attenuation-contamination formula ($\Lambda = \Sigma_{\mathbf{T}}\Sigma_{\mathbf{Q}}^{-1}$), we express the elements of Λ as follows

$$\lambda_{11} = (\,\sigma_{T_1}^2\sigma_{Q_2}^2 - \sigma_{T_1T_2}\sigma_{Q_1Q_2})/(\sigma_{Q_1}^2\sigma_{Q_2}^2 - \sigma_{Q_1Q_2}.\sigma_{Q_1Q_2}),$$
$$\lambda_{12} = (\sigma_{Q_1}^2\sigma_{T_1T_2} - \sigma_{T_1}^2\sigma_{Q_1Q_2})/(\sigma_{Q_1}^2\sigma_{Q_2}^2 - \sigma_{Q_1Q_2}.\sigma_{Q_1Q_2}),$$
$$\lambda_{21} = (\sigma_{Q_2}^2\sigma_{T_1T_2} - \sigma_{T_2}^2\sigma_{Q_1Q_2})/(\sigma_{Q_1}^2\sigma_{Q_2}^2 - \sigma_{Q_1Q_2}.\sigma_{Q_1Q_2}),$$
$$\lambda_{22} = (\sigma_{T_2}^2\sigma_{Q_1}^2 - \sigma_{T_1T_2}\sigma_{Q_1Q_2})/(\sigma_{Q_1}^2\sigma_{Q_2}^2 - \sigma_{Q_1Q_2}.\sigma_{Q_1Q_2}).$$

Further, if the true association parameters for the exposure and confounder are β_{T_1} and β_{T_2}, respectively, then the observed associations when DQ data are used would be

$$\begin{aligned} \beta_{Q_1} &= \lambda_{11}\beta_{T_1} + \lambda_{12}\beta_{T_2}, \\ \beta_{Q_2} &= \lambda_{22}\beta_{T_2} + \lambda_{21}\beta_{T_1}, \end{aligned} \tag{6.5}$$

respectively. The interest is mainly in the exposure effect β_{T_1}. The term $\lambda_{11}\beta_{T_1}$ is the attenuation expression as for a single exposure variable measured with error. The second part of the expression $\lambda_{12}\beta_{T_2}$ is the contamination expression introduced by the confounder. If the contamination factor λ_{12} and the confounder effect β_{T_2} are large, the bias in β_{T_1} will be substantial. For instance, if there is a strong positive correlation between the confounder and the exposure such that $\lambda_{12}, \lambda_{21} > 0$ and positive associations between the two variables with the outcome ($\beta_{T_1}, \beta_{T_2} > 0$), the observed association will be inflated ($\beta_{Q_1} > \beta_{T_1}$). In contrast, if there is weak positive association between the exposure and the outcome, and a strong negative association between the confounder and the outcome such that $\lambda_{11}\beta_{T_1} + \lambda_{12}\beta_{T_2} < 0$, the direction of the association will be reversed. However, if λ_{12} is close to zero, i.e., the confounder is weakly correlated with the exposure, the observed association will mostly be attenuated with λ_{11}. In other words, if the confounder is weakly linked with the outcome and exposure variables, the effect of confounding on the exposure-outcome association will not be substantial, and the univariate adjustment method for measurement error will be sufficient. Now assume that the exposure is not associated with the outcome (i.e., $\beta_{T_1} = 0$), and that the confounder is strongly linked with the exposure and the outcome, then the exposure will appear to be associated with the outcome (i.e., $\beta_{Q_1} = \lambda_{12}\beta_{T_2} \neq 0$). Finally, suppose that the two variables are correlated through their true values, and that the exposure is measured exactly, whereas the confounder is measured with error, the observed exposure-outcome association will still be biased. The bias is due to contamination by correlation between their true values (i.e., $\lambda_{12} \neq 0$). From the above illustration, some of the serious implications of the confounder measurement error are as follows: (i) the health risk associated with a given risk factor might be underestimated, (ii) the health benefit attributable to increased intake of a given exposure, say, fresh vegetables, might be severely underestimated, and that (iii) a risk factor may appear beneficial to health, when indeed the opposite is the truth. The implications of confounder measurement error can be enormous (Kupper, 1984). This, therefore, calls for thorough scrutiny while interpreting study findings; for instance, by using the method proposed in chapter 4.

6.2.3 Limitations, and their implications

6.2.3.1 Information on error correlation

An important limitation of our study is that it requires information on the correlation of the error in the exposure and the error in the confounder. In practice this information is seldom available. We proposed to use sensitivity analysis to try a range of sensible values in order determine the possible effects of measurement error. This could be used to turn around the problem such that instead of using the method to estimate the true association, one could use the method to determine how large the error correlation should be in order to fully explain the observed crude association between an exposure and a health outcome.

6.2.3.2 Validity coefficients using concentration markers and dietary records/recalls

In eliciting information on validity measures of self-report instruments, we used literature data on the correlation coefficient between true and observed intakes (validity coefficient). The literature data on validity coefficients were derived from studies, where concentration markers and dietary recalls/records were used as reference instruments. However, these instruments do not provide direct measures of true intake (Andersen et al., 2005; Slater et al., 2010). Therefore, the validity coefficients cannot be determined exactly, due to the bias in these instruments. To illustrate this using the measurement error model (6.1) and the validity coefficient formula (Kipnis et al., 2003), the validity coefficient ($\rho_{R,Q}$) obtained with biased concentration markers or dietary recalls does not equal the true validity coefficient ($\rho_{T,Q}$). The two validity coefficients are related as

$$\rho_{R,Q} = \rho_{T,Q} \frac{\sigma_T}{\sigma_R} \left(\beta_R + \frac{1}{\beta_Q} \rho_{qr} \sqrt{\frac{\sigma_q^2}{\sigma_T^2} \frac{\sigma_r^2}{\sigma_T^2}} \right). \qquad (6.6)$$

When there is intake related bias (β_R) and correlated person-specific bias (ρ_{qr}) in these instruments, the validity of the observed data will misrepresent the true validity. As a consequence, the estimated associations, adjusted with the method proposed in chapter 4, might still be biased.

6.2.3.3 Transportability of the external validation data

Another issue that requires keen assessment is whether validity coefficients and attenuation factors obtained from the literature studies (in chapter 4) are transportable to the primary study; that is, whether these literature data can be transferred to the primary study without causing bias in estimating exposure-outcome associations (Carroll et al., 2006). If the primary-study population is different from the populations where the external validation data were obtained, the external information will not be transportable to the primary study and the exposure-outcome association would be prone to estimation bias. A simple way to ensure transportability could be to include external validity data only from studies with roughly the same distribution of subject characteristics (e.g., age group, BMI etc.), and using similar self-report instruments as in the primary study. Nevertheless, even if the external information is not transportable, our method gives an indication of the effect of measurement error.

6.2.4 Possible extensions and applications of the model

Self-reports may depend also on subject characteristics such as body mass index and age. To accommodate these additional systematic effects, the measurement model shown in chapter 4 can be extended by adding systematic effect terms for these subject characteristic variables. However, this extension is only possible if information on these additional variables and prior information on the magnitude of their effects is available in the main study.

Additionally, it might be insightful to assess the effects of the unmeasured confounders. In a large study, it might not be possible for an investigator to measure all the confounders. The confounders might be missed for reasons of cost, by happenstance or due to flaws in the data collection process. The observed exposure-outcome association that is adjusted for exposure measurement error, confounding by other observed variables and measurement error in the observed confounders will be a mixture of the effect of the exposure and unobserved confounders (Gustafson and McCandless, 2010). The method proposed in chapter 4 can be modified and used in a sensitivity analysis to assess the effect of the unmeasured confounders on the adjusted exposure-outcome associations. This can be done by including the components for the unmeasured confounder in place of the measured confounder. First, because the confounder is unmeasured, its association with the outcome is assumed as the true association; thus in expression (6.5), $\beta_{Q_2} = \beta_{T_2}$. Second, in the same expression and because the confounder is unmeasured, measurement error components for the unmeasured confounder are set to zero, such that $\sigma^2_{Q_2} = \sigma^2_{T_2}$, and $\sigma_{Q_1Q_2} = \sigma_{Q_1T_2} = \sigma_{T_1T_2}$ under the assumption that measurement errors in the exposure are uncorrelated with the true unmeasured confounder (i.e., $\sigma_{\epsilon_{Q_1},T_2} = 0$). These modifications can be used in expression (4.3) in chapter 4 to assess the effects of unmeasured confounders. Thus, the effect of the unmeasured confounder on the exposure-outcome association can be assessed in a sensitivity analysis. The sensitivity analysis can be done with respect to: (i) the strength of association between the unmeasured confounder and the outcome and (ii) the magnitude of correlation between the exposure and the unmeasured confounder, and (iii) how the estimate of true exposure-outcome association depends on the unmeasured confounder.

For illustration, we use mortality (outcome), vegetable intake (exposure) and alcohol intake (unmeasured confounder) variables as an example. Using this example, the first part of the sensitivity analysis can be done by varying the assumed association between the unmeasured alcohol intake and mortality, while assuming a fixed value for the correlation between true vegetable intake and true alcohol intake. The second part can be done by varying the magnitude of the assumed correlation between true vegetable intake and true alcohol intake, while assuming a fixed value for the association between the unmeasured confounder and mortality. The last part of the sensitivity analysis can be done by assessing the discrepancy between the estimated true association between vegetable intake and mortality adjusted for the assumed effect of alcohol intake (i.e., using the modified multivariate bias-adjustment method) and the unconditional association between vegetable intake and mortality not adjusted for the assumed effect of alcohol intake (i.e., using the univariate method shown in chapter 4).

6.3 Measurement error modelling for physical activity in free-living individuals as measured by an accelerometer

6.3.1 Main findings

Assessment of physical activity with accelerometry is an area of active research. We fitted a measurement error model on data collected with a GT3X accelerometer (chapter 5). First, we observed that the accelerometer underestimated an individual's daily level of physical activity. For instance, the mean of physical activity level was underestimated by about 0.3. Second, because of measurement error the association between physical activity and an outcome would be attenuated. With an attenuation factor of 0.55 for physical activity level (see chapter 5), a true relative risk of 0.5 for protective effect of regular physical activity would appear approximately as 0.7, i.e., weaker than the true association. Therefore, health benefits associated with regular physical activity would be underestimated, when physical activity data from the accelerometer are used to estimate the association between physical activity and health outcome. Lastly, we found that due to measurement error in the accelerometer, there would be loss of statistical power to detect significant associations. This means that the size of the accelerometer study would need to be very large in order to detect existing associations. For instance, with a correlation coefficient of 0.44 between true and measured physical activity level (chapter 5), the sample size of the accelerometer study would need be about five times (0.44^{-2}) as large as the sample size that would be required if physical activity were measured exactly.

6.3.2 Limitations

In the physical activity study presented in chapter 5, the study participants were all adults, of the same ethnicity and from the same region. Thus, it might be misleading to generalize the findings from that study to other populations. A fundamental aspect of a good study is its external validity. External validity refers to the ability to generalize the analysis results from study population (e.g., those in chapter 5) to other populations, i.e., extending inferences about a source population to a target population (Rothman et al., 2008, p. 128). A potential threat to the external validity might be that the study sample is not a representative of the study population. To ensure generalizability, a study with a more diverse sample can be performed by including subjects who are more representative of the whole population. However, with a more diverse population, there is an increased likelihood of confounding by other factors. Therefore, it is advantageous to restrict studies to subjects with comparable characteristics and on whom complete and precise information can be obtained.

6.3.3 Implications

Similar to 24HR and example in section 6.1.2.2, use of accelerometer as a “reference” instrument to validate other instrument, such as physical activity questionnaires, may underestimate the true effects of measurement error in the latter (Arem, Keadle and Matthews, 2015).

6.3.4 Possible extensions of the model

The measurement error model presented in chapter 5 can be extended to include the systematic effects in the accelerometer that are contributed by subject characteristic variables such as, level of education and age, as explained in 6.2.4.

6.3.5 A comparison of the chosen modelling method to Bayesian methods

Although maximum likelihood was used to model measurement error in physical activity, Bayesian methods could also be used (Dellaportas and Stephens, 1995; Richardson and Gilks, 1993). With the Bayesian approach, prior information on exposure measurement error can be incorporated (Apanasovich, Carroll and Maity, 2009; Carroll et al., 2006). In addition, Bayesian sampling-based methods, such as Markov Chain Monte Carlo (MCMC), may provide more flexibility to include other measurement error structures. Moreover, Bayesian MCMC can be implemented in many generic software such as WinBUGS (Lunn et al., 2000), Open Bugs (Lunn et al., 2009) and JAGS (Plummer, 2003). Additionally, non-sampling based integrated nested Laplace approximation (INLA) (Rue et al., 2009) might be more efficient in measurement error modelling (Muff, 2015).

To illustrate the use of Bayesian methods, we reanalysed physical activity data presented in chapter 5 using Bayesian MCMC and INLA and compared the results with those from the maximum likelihood method. The task was to quantify measurement error in physical activity (expressed as total energy expenditure) as measured by a GT3X accelerometer using a validity coefficient to quantify loss of statistical power, and an attenuation factor to quantify bias in the association between physical activity and a health outcome. Non-informative priors were used for the model parameters and hyperparameters. The MCMC model was specified as follows: three chains were used and for each chain, 100 000 burn-in samples were discarded and every 5th of the remaining 500 000 samples were retained. In total, MCMC analysis was based on 300 000 samples. For comparison purposes, a frequentist maximum likelihood estimation (MLE) method with adaptive quadrature was used, and 95% confidence interval (CI) was estimated by bootstrapping (Agogo, 2015). Figure 6.1 presents the posterior distributions (and credible intervals) from the Bayesian analyses and the point estimate (and confidence interval) from the MLE analysis for the validity coefficient and the attenuation factor. The results from the three estimation methods are quite similar, with INLA having the fastest computation time. The computation times for the three methods were as follows: MCMC (11.19 minutes), MLE (20.07 seconds), and INLA (2.87 seconds). Despite the flexibility of Bayesian methods, MCMC can be computationally expensive for complex measurement error models.

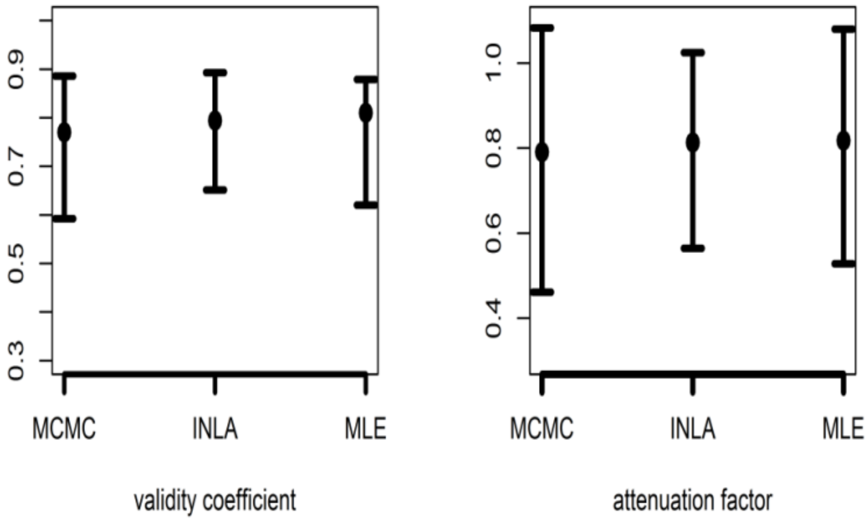

Figure 6.1: Mean estimates and 95%CI for validity coefficient and attenuation factor with Bayesian MCMC, INLA and frequentist MLE (right panel); CI=credible interval for MCMC and INLA, CI=confidence interval for MLE.

6.4 Measurement error correction methods for categorized exposures

In this thesis, we focussed on measurement error in continuous exposure variables. However, in epidemiologic studies, an exposure variable is often categorized into quantiles and exposure-outcome association estimated within each category relative to the reference category (Keogh and White, 2014). In chapter 2, the focus could have been on adjusting for measurement error in vegetable subgroup intakes categorized into quantiles instead of on a continuous scale. Measurement error in a continuous exposure may lead to misclassification when the exposure is categorized, leading to biased estimates of relative association between the exposure categories and outcome. If the task involves adjusting for measurement error in a categorized exposure using a single-replicate validation study, graphical method of Macmahon et al. (1990) and regression calibration can be used in a linear or approximately linear exposure-outcome model. In the graphical method, the crude odds ratio or hazard ratio (that is, not adjusted for measurement error) are plotted against the estimated true mean intake in each category. In this case, the true mean intake is estimated as the mean of the reference measurement for the subjects in each category (Macmahon et al., 1990). However, this method assumes classical measurement error, does not usually work for non-linear associations and does not account for multivariate measurement error (Keogh and White, 2014). In the regression calibration method for categorized exposures, the calibration is done on the ranking of the quantiles rather than on the intake value themselves (as in the continuous case) under the assumption of a linear trend as described in Keogh and White (2014).

6.5 Suggestions for future research

The assumptions of a valid reference instrument highlighted in section 6.1.2 can be assessed in a simulation study. This can be done by extending the simulation study presented in chapter 3. The extension can be done by including intake-related bias terms in the simulation model for the 24HR intake and introducing positive correlation between the random errors in the 24HR and random errors in the DQ intake, using realistic values from validation studies such as the OPEN study (Subar et al., 2002). This type of simulation study, however, is hampered because not many valid reference instruments have been identified for dietary intakes. Only very few dietary biomarkers qualify as reference instruments for dietary intake, as most existing biomarkers only measure concentrations for which the quantitative relation to dietary intake is unknown (Kaaks, 1997). Such concentration biomarkers, therefore, cannot be used as valid reference measurements and can only be used as correlates of intake (Kipnis et al., 2003). Thus, a new technology of monitoring dietary intake behaviour using sensors and/or video registration might provide a better improvement and can be an area for future research. Presently, further work could focus on exploring whether dietary data from self-report instruments and concentration biomarkers can be combined to assess dietary exposures with no reference biomarkers.

The work in this thesis shows the need for better empirical evidence on the correlation between measurement errors in the confounders and exposures. Although validation studies have been carried out, for instance, the OPEN study (Subar et al., 2003), we are not aware of any analysis on the correlation between errors of important confounders such as smoking and dietary intake presented in chapter 5. Such work is needed in order to better understand the findings in nutritional epidemiology.

6.6 Conclusion

The problem of exposure measurement error on study findings has attracted a lot of methodological research in the recent past. However, application of most available methods requires validation studies with multiple replicate data, whereas some studies such as EPIC have only a single replicate measurement per individual. Moreover, in large epidemiological studies, it might not be feasible to conduct internal validation studies. In such an epidemiological study without an internal validation study, it is common to have potential confounders that are often mismeasured. Additionally, some "reference" instruments being used to validate other error-prone instruments are themselves marred by measurement error, which can lead to erroneous validity measures for the instruments being validated.

Related to these three issues, the following questions were the main motivation for the work in this thesis:

(i) How to adjust for exposure measurement error using a single-replicate validation study with zero-inflated reference measurements.

(ii) How to adjust the exposure-outcome association for exposure measurement error, confounding and measurement error in the confounders when there is no internal validation study.

(iii) How to quantify measurement error in physical activity as measured by accelerometer in a recently concluded validation study.

From this thesis, the following can be concluded:

(i) A suitably specified two-part regression calibration can be used to adjust for the bias in the exposure-outcome associations using a single-replicate validation study with zero-inflated reference measurements, although a suitable specified one-part regression calibration also performs quite well.

(ii) In the absence of an internal validation study, carefully extracted data on validity of self-report instruments can be used to adjust for the bias in the exposure-outcome associations. The external validity data, however, should be transferable to the main study. The method proposed in this thesis is useful in understanding the effects of error correlations on the exposure-outcome associations.

(iii) When an accelerometer (such as GT3X) is used to validate other instruments in measuring physical activity, the effect of measurement error can be seriously underestimated, because accelerometers are marred by substantial measurement error. As a result, the study conclusions might be misleading.

Glossary

Accelerometer: is a motion sensor that consists of piezoelectric transmitters used for monitoring body accelerations. With the accelerometer, the frequency, intensity, and duration of physical activity can be assessed as a function of body movement.

Activity Energy Expenditure (AEE): refers to energy expended above resting level due to physical activity.

Attenuation: when measurement errors in the exposure variable bias a regression coefficient that quantifies the exposure-outcome association towards zero, we refer to bias of that nature as attenuation

Attenuation factor: is a measure of the amount of attenuation in a relation due to exposure measurement error. For a linear relation between measured and true exposure, the attenuation factor is equal to the regression slope of true on measured exposure. An attenuation factor close to 1 indicates minimum attenuation, whereas a factor close to 0 indicates maximum attenuation.

Attenuation-contamination matrix: extension of attenuation factor to multivariate exposures. The diagonal elements are called attenuation factors and off-diagonals elements are called contamination factors.

Basal Energy Expenditure (BEE): refers to the energy expended at rest in a fasting state.

Contamination factor: determines how much the confounder contributes to the bias of the estimated exposure-outcome association. A contamination factor close to zero indicates minimum contamination, and that contamination increases with the absolute magnitude of the contamination factor.

Calibration/validation study: a study where next to the main-study instrument also unbiased exposure measurements are obtained. The validation study is often used to determine the validity of the main-study instrument.

Dietary Questionnaire (DQ): is a type of questionnaire for assessing long-term dietary intake where individuals report their past intakes (usually from several months to a year) of various dietary or groups of dietary components.

Doubly Labelled Water (DLW): is a technique used to assess total energy expenditure in a free-living context.

Food Frequency Questionnaire (FFQ): the food frequency questionnaire consists of a list of foods and a selection of options relating to the frequency of consumption of each of the foods listed (e.g. times per day, daily, weekly, and monthly). FFQs are designed to collect dietary information from large numbers of individuals and are normally self-administered, though interviewer administered and telephone interview are possible. FFQs normally ask about intake within a given time frame (e.g. in the past 2-3 months, 1 year or longer) and therefore aim to capture long-term average intake. The length of the food list can vary depending on the nutrients or foods of interest.

Intake related/proportional scaling bias: refers to the bias in, say, dietary questionnaire, that is related with the level of intake. A proportional scaling bias factor close to zero indicates severe underestimation of intake for an individual with high level of true intake and a value greater than 1 indicates overestimation of true intake.

Measurement error: refers to the discrepancy between the true and measured values of a variable.

Nondifferential error: occurs when the measured exposure contains no extra information about the health outcome over what is contained in the true exposure.

Person-specific bias: is the component of bias that describes the fact that two individuals who consume the same amount of food will systematically report their intakes differently.

Physical activity (PA): refers to bodily movements due to contraction of skeletal muscle that result in an increase in energy expenditure above resting levels.

Physical activity level (PAL): refers to the ratio of total energy expenditure to basal energy expenditure and provides an index of the average relative excess output related to physical activity.

Random measurement error: random error arises when the measured exposure is distributed randomly around the true value; in this case the exposure is sometimes overestimates and sometimes underestimates true exposure level.

Reference measurement: refers to the unbiased measurements in a validation/calibration study.

Regression calibration: Regression calibration involves replacing the observed exposure value with the conditional expectation of true exposure given the observed data in estimating the exposure-outcome association.

Systematic bias: systematic bias arises when an individual consistently overestimates or underestimates his true level of exposure, such that the average of measured exposure does not equal the true mean.

Total Energy Expenditure (TEE): is the sum of energy expended at rest (BEE), energy expended above resting level due to physical activity (AEE) and the thermic effect of food.

Twenty-four hour recall (24HR): in the 24HR, an individual is asked to report dietary intake for the past 24 hours. Usually, the recall is conducted by personal interview. As a retrospective method, 24HR relies on an accurate memory of intake, reliability of the respondent not to under/misreport, and an ability to estimate portion size.

Validity coefficient: is the correlation coefficient between measured and true exposure values; the validity coefficient can be used to quantify the loss of statistical power to detect associations. A value close to zero indicates severe loss of statistical power to detect associations, whereas a value close to one indicates minimal loss of statistical power.

References

Agogo, G. O., van der Voet, H., van' t Veer, P.., *et al.* (2014). Use of Two-Part Regression Calibration Model to Correct for Measurement Error in Episodically Consumed Foods in a Single-Replicate Study Design: EPIC Case Study. *PLoS ONE* **9**, e113160.

Agogo, G. O., van der Voet, H.,Trijsburg, L., van Eeuwijk, F.A., van 't Veer, P., Boshuizen, H.C. (2015). Measurement error modelling for accelerometer activity data using Bayesian integrated nested Laplace approximation. In *International Workshop on Statistical Modelling*, H. F. H. Wagner (ed), 3-6. Linz, Austria.

Agudo, A. (2004). Measuring intake of fruit and vegetables. In *Background paper for the Joint FAO/WHO Workshop on Fruits and Vegetables*. Kobe,Japan: WHO.

Agudo, A., Cabrera, L., Amiano, P., *et al.* (2007). Fruit and vegetable intakes, dietary antioxidant nutrients, and total mortality in Spanish adults: findings from the Spanish cohort of the European Prospective Investigation into Cancer and Nutrition (EPIC-Spain). *American Journal of Clinical Nutrition* **85**, 1634-1642.

Al-Delaimy, W. K., Ferrari, P., Slimani, N., *et al.* (2005). Plasma carotenoids as biomarkers of intake of fruits and vegetables: individual-level correlations in the European Prospective Investigation into Cancer and Nutrition (EPIC). *European Journal of Clinical Nutrition* **59**, 1387-1396.

Andersen, L. F., Veierod, M. B., Johansson, L., Sakhi, A., Solvoll, K., and Drevon, C. A. (2005). Evaluation of three dietary assessment methods and serum biomarkers as measures of fruit and vegetable intake, using the method of triads. *British Journal of Nutrition* **93**, 519-527.

Apanasovich, T. V., Carroll, R. J., and Maity, A. (2009). SIMEX and standard error estimation in semiparametric measurement error models. *Electron J Stat* **3**, 318-348.

Arem, H., Keadle, S. K., and Matthews, C. E. (2015). Invited commentary: meta-physical activity and the search for the truth. *American Journal of Epidemiology* **181**, 656-658.

Bateson, T. F., and Wright, J. M. (2010). Regression Calibration for Classical Exposure Measurement Error in Environmental Epidemiology Studies Using Multiple Local Surrogate Exposures. *American Journal of Epidemiology* **172**, 344-352.

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine* **24**, 1713-1723.

Berkson, J. (1950). Are there two regressions? *Journal of the American Statistical Association* **45**, 164-180.

Bingham, S. A., Gill, C., Welch, A., *et al.* (1997). Validation of dietary assessment methods in the UK arm of EPIC using weighed records, and 24-hour urinary nitrogen and potassium and serum vitamin C and carotenoids as biomarkers. *International Journal of Epidemiology* **26 Suppl 1**, S137-151.

Bingham, S. A., Gill, C., Welch, A., *et al.* (1994). Comparison of dietary assessment methods in nutritional epidemiology: weighed records v. 24 h recalls, food-frequency questionnaires and estimated-diet records. *Br J Nutr* **72**, 619-643.

Black, A. E., Welch, A. A., and Bingham, S. A. (2000). Validation of dietary intakes measured by diet history against 24 h urinary nitrogen excretion and energy expenditure measuredby the doubly-labelled water method in middle-aged women. *British Journal of Nutrition* **83**, 341-354.

Bland, J. M., and Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **1**, 307-310.

Block, G. (1982). A review of validations of dietary assessment methods. *American Journal of Epidemiology* **115**, 492-505.

Block, G. (1989). Human dietary assessment: methods and issues. *Preventive Medicine* **18**, 653-660.

Boffetta, P., Couto, E., Wichmann, J., *et al.* (2010). Fruit and Vegetable Intake and Overall Cancer Risk in the European Prospective Investigation Into Cancer and Nutrition (EPIC). *Journal of the National Cancer Institute* **102**, 529-537.

Boshuizen, H. C., Lanti, M., Menotti, A., *et al.* (2007). Effects of past and recent blood pressure and cholesterol level on coronary heart disease and stroke mortality, accounting for measurement error. *American Journal of Epidemiology* **165**, 398-409.

Buonaccorsi, J. P. (2010). Measurement error : models, methods, and applications. Boca Raton: CRC Press.

Burton, A., Altman, D., Royston, P., and Holder, R. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, 4279-4292.

Cai, W. (2008). Fitting Generalized Additive Models with the GAM Procedure in SAS 9.2. In *SAS Global Forum 2008*. SAS Campus Drive, Cary,NC 27513: SAS Institute Inc.

Carroll, R. J., Delaigle, A., and Hall, P. (2009). Nonparametric Prediction in Measurement Error Models. *Journal of the American Statistical Association* **104**, 993-1014.

Carroll, R. J., Freedman, L. S., and Kipnis, V. (1998). Measurement error and dietary intake. *Adv Exp Med Biol* **445**, 139-145.

Carroll, R. J., Kuchenhoff, H., Lombard, F., and Stefanski, L. A. (1996). Asymptotics for the SIMEX estimator in nonlinear measurement error models. *Journal of the American Statistical Association* **91**, 242-250.

Carroll, R. J., Midthune, D., Subar, A. F., *et al.* (2012). Taking Advantage of the Strengths of 2 Different Dietary Assessment Instruments to Improve Intake Estimates for Nutritional Epidemiology. *American Journal of Epidemiology* **175**, 340-347.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models*. New York: Chapman& Hall/CRC.

Cassell, D. L. (2007). Don't Be Loopy: Re-Sampling and Simulation the SAS® Way. In *SAS Global Forum : Statistics and Data Analysis*. Corvallis, Orlando, Florida.

Chumlea, W. C., Guo, S. S., Kuczmarski, R. J., *et al.* (2002). Body composition estimates from NHANES III bioelectrical impedance data. *International Journal of Obesity* **26**, 1596-1609.

Cole, S. R., Chu, H. T., and Greenland, S. (2006). Multiple-imputation for measurement-error correction. *International Journal of Epidemiology* **35**, 1074-1081.

Cook, J. R., and Stefanski, L. A. (1994). Simulation-Extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association* **89**, 1314-1328.

Cox, D. R. (1970). *The analysis of binary data*. London: Methuen.

Cox, D. R. (1972). Regression Models and Life-Tables. Journal of the Royal Statistical Society Series B-Statistical Methodology **34**, 187-220.

Crouter, S. E., Churilla, J. R., and Bassett, D. R., Jr. (2006). Estimating energy expenditure using accelerometers. *European Journal of Applied Physiology* **98**, 601-612.

Day, N. E., McKeown, N., Wong, M. Y., Welch, A., and Bingham, S. (2001). 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**, 309-317.

Day, N. E., Wong, M. Y., Bingham, S., *et al.* (2004). Correlated measurement error - implications for nutritional epidemiology. *International Journal of Epidemiology* **33**, 1373-1381.

de Boer, W. J., van der Voet, H. (2010). A web-based program for Monte Carlo Risk Assessment (MCRA). Netherlands: Biometris-Wageningen UR, RIKILT, Institute of Food Safety-Wageningen UR and National Institute for Public Health and the Environment (RIVM).

Dellaportas, P., and Stephens, D. A. (1995). Bayesian-analysis of errors-in-variables regression-models. *Biometrics* **51**, 1085-1095.

Efron, B., and Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Eliopoulos, C., Klein, J., and Koren, G. (1996). Validation of self-reported smoking by analysis of hair for nicotine and cotinine. *Therapeutic Drug Monitoring* **18**, 532-536.

Ferrari, P., Day, N. E., Boshuizen, H. C., *et al.* (2008). The evaluation of the diet/disease relation in the EPIC study: considerations for the calibration and the disease models. *International Journal of Epidemiology* **37**, 368-378.

Ferrari, P., Friedenreich, C., and Matthews, C. E. (2007). The role of measurement error in estimating levels of physical activity. *American Journal of Epidemiology* **166**, 832-840.

Ferrari, P., Kaaks, R., Fahey, M. T., *et al.* (2004). Within- and between-cohort variation in measured macronutrient intakes, taking account of measurement errors, in the European Prospective Investigation into cancer and nutrition study. *American Journal of Epidemiology* **160**, 814-822.

Feskanich, D., Rimm, E. B., Giovannucci, E. L., *et al.* (1993). Reproducibility and validity of food-intake measurements from a semiquantitative food frequency questionnaire. *Journal of the American Dietetic Association* **93**, 790-796.

Fraser, G. E., and Stram, D. O. (2001). Regression calibration in studies with correlated variables measured with error. *American Journal of Epidemiology* **154**, 836-844.

Fraser, G. E., and Stram, D. O. (2012). Regression calibration when foods (measured with error) are the variables of interest: markedly non-Gaussian data with many zeroes. *American Journal of Epidemiology* **175**, 325-331.

Freedman, L. S., Fainberg, V., Kipnis, V., Midthune, D., and Carroll, R. J. (2004). A new method for dealing with measurement error in explanatory variables of regression models. *Biometrics* **60**, 172-181.

Freedman, L. S., Midthune, D., Carroll, R. J., and Kipnis, V. (2008). A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression. *Statistics in Medicine* **27**, 5195-5216.

Freedman, L. S., Schatzkin, A., Midthune, D., and Kipnis, V. (2011). Dealing With Dietary Measurement Error in Nutritional Cohort Studies. *Journal of the National Cancer Institute* **103**, 1086-1092.

Freedson, P. S., Melanson, E., and Sirard, J. (1998). Calibration of the Computer Science and Applications, Inc. accelerometer. *Medicine and science in sports and exercise* **30**, 777-781.

Fuller, W. A. (2006). *Measurement error models*. Hoboken, N.J.: Wiley-Interscience.

Geert Molenberghs, M. G. K. (2007). *Missing Data in Clinical Studies*. West Sussex: John Wiley & Sons.

Gelman, A., and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge ; New York: Cambridge University Press.

Goedhart, P. W., van der Voet, H., Knuppel, S., Dekkers, A.L.M., Boeing, H.,Dodd, K.W., van Klaveren, J. (2012). A comparison by simulation of different methods to estimate the usual intake distribution for episodically consumed foods. *Supporting Publications 2012:EN-299* **66**.

Goldbohm, R. A., Vandenbrandt, P. A., Brants, H. A. M., *et al.* (1994). Validation of a dietary questionnaire used in a large-scale prospective cohort study on diet and cancer. *European Journal of Clinical Nutrition* **48**, 253-265.

Gorsuch, R. L., Lehmann, C.S (2010). Correlation Coefficients: Mean Bias and Confidence Interval Distortions. *Journal of Methods and Measurement in the Social Sciences* **1**, 52-65.

Greenland, S. (1989). Modeling and variable selection in epidemiologic analysis. *American Journal of Public Health* **79**, 340-349.

Guidotti, S., Jansen, H. G., Aerts-Bijma, A. T., Verstappen-Dumoulin, B. M. A. A., Van Dijk, G., and Meijer, H. A. J. (2013). Doubly Labelled Water analysis: Preparation, memory correction, calibration and quality assurance for δ2H and δ18O measurements over four orders of magnitudes. **27**, 1055-1066.

Guolo, A. (2008). A Flexible Approach to Measurement Error Correction in Case-Control Studies. *Biometrics* **64**, 1207-1214.

Guolo, A., and Brazzale, A. R. (2008). A simulation-based comparison of techniques to correct for measurement error in matched case-control studies. *Statistics in Medicine* **27**, 3755-3775.

Gustafson, P., and McCandless, L. C. (2010). Probabilistic Approaches to Better Quantifying the Results of Epidemiologic Studies. *International Journal of Environmental Research and Public Health* **7**, 1520-1539.

Hallal, P. C., Reichert, F. F., Clark, V. L., *et al.* (2013). Energy Expenditure Compared to Physical Activity Measured by Accelerometry and Self-Report in Adolescents: A Validation Study. *PLoS ONE* **8**.

Hastie, T., and Tibshirani, R. (1999). *Generalized additive models*. Boca Raton, Fla.: Chapman & Hall/CRC.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The elements of statistical learning : data mining, inference, and prediction*, 2nd edition. New York, NY: Springer.

Henry, C. J. (2005). Basal metabolic rate studies in humans: measurement and development of new equations. *Public Health Nutrition* **8**, 1133-1152.

Hills, A. P., Mokhtar, N., and Byrne, N. M. (2014). Assessment of physical activity and energy expenditure: an overview of objective measures. *Frontiers in Nutrition* **1**, 5.

Huang, Y., Chen, R., and Dagne, G. (2011). Simultaneous Bayesian Inference for Linear, Nonlinear and Semiparametric Mixed-Effects Models with Skew-Normality and Measurement Errors in Covariates. *International Journal of biostatistics* **7**.

IAEA (2009). Assessment of Body Composition and Total Energy Expenditure in Humans Using Stable Isotope Techniques. Vienna: International Atomic Energy Agency

Johnson, R. K. (2002). Dietary intake - How do we measure what people are really eating? *Obesity Research* **10**, 63S-68S.

Kaaks, R. (1997). Biochemical markers as additional measurements in studies of the accuracy of dietary questionnaire measurements:conceptual issues. *American Journal of Clinical Nutrition*, 1232s-1239s.

Kaaks, R., Ferrari, P., Ciampi, A., Plummer, M., and Riboli, E. (2002). Uses and limitations of statistical accounting for random error correlations, in the validation of dietary questionnaire assessments. *Public Health Nutrition* **5**, 969-676.

Kaaks, R., and Riboli, E. (1997). Validation and calibration of dietary intake measurements in the EPIC project: Methodological considerations. *International Journal of Epidemiology* **26**, S15-S25.

Kaaks, R., Slimani, N., and Riboli, E. (1997). Pilot phase studies on the accuracy of dietary intake measurements in the EPIC project: Overall evaluation of results. *International Journal of Epidemiology* **26**, S26-S36.

Keogh, R. H., and White, I. R. (2014). A toolkit for measurement error correction, with a focus on nutritional epidemiology. *Statistics in Medicine* **33**, 2137-2155.

Keogh, R. H., White, I. R., and Rodwell, S. A. (2013). Using surrogate biomarkers to improve measurement error models in nutritional epidemiology. *Statistics in Medicine* **32**, 3838-3861.

Kipnis, V., Carroll, R. J., Freedman, L. S., and Li, L. (1999). Implications of a new dietary measurement error model for estimation of relative risk: application to four calibration studies. *American Journal of Epidemiology* **150**, 642-651.

Kipnis, V., Freedman, L., Brown, C., Hartman, A., and Schatzkin, A. (1997). Effect of Measurement Error on Energy-Adjustment Models in Nutritional Epidemiology. **146**, 842-854.

Kipnis, V., Freedman, L. S., Carroll, R. J., and Midthune, D. (2015). A bivariate measurement error model for semicontinuous and continuous variables: Application to nutritional epidemiology. *Biometrics*.

Kipnis, V., Midthune, D., Buckman, D. W., *et al.* (2009). Modeling data with excess zeros and measurement error: application to evaluating relationships between episodically consumed foods and health outcomes. *Biometrics* **65**, 1003-1010.

Kipnis, V., Midthune, D., Freedman, L. S., *et al.* (2001). Empirical evidence of correlated biases in dietary assessment instruments and its implications. *American Journal of Epidemiology* **153**, 394-403.

Kipnis, V., Subar, A. F., Midthune, D., *et al.* (2003). Structure of dietary measurement error: results of the OPEN biomarker study. *American Journal of Epidemiology* **158**, 14-21; discussion 22-16.

Krouwer, J. S. (2008). Why Bland-Altman plots should use X, not (Y+X)/2 when X is a reference method. *Statistics in Medicine* **27**, 778-780.

Kupper, L. L. (1984). Effects of the use of unreliable surrogate variables on the validity of epidemiologic research studies. *American Journal of Epidemiology* **120**, 643-648.

Leenders, N. Y. J. M., Sherman, W. M., Nagaraja, H. N., and Kien, C. L. (2001). Evaluation of methods to assess physical activity in free-living conditions. *Medicine and science in sports and exercise* **33**, 1233-1240.

Lesaffre, E., and Lawson, A. (2012). *Bayesian biostatistics*. Chichester, West Sussex: John Wiley & Sons.

Lim, S., Wyker, B., Bartley, K., and Eisenhower, D. (2015). Measurement error of self-reported physical activity levels in new york city: assessment and correction. *American Journal of Epidemiology* **181**, 648-655.

Lin, X. H., and Carroll, R. J. (1999). SIMEX variance component tests in generalized linear mixed measurement error models. *Biometrics* **55**, 613-619.

Lu, Z. (2006). Computation of Correlation Coefficient and Its Confidence Interval in SAS. In *SUGI 31 Proceedings*. San Francisco, California

Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine* **28**, 3049-3067.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing* **10**, 325-337.

Lyden, K., Kozey, S. L., Staudenmeyer, J. W., and Freedson, P. S. (2011). A comprehensive evaluation of commonly used accelerometer energy expenditure and MET prediction equations. *European Journal of Applied Physiology* **111**, 187-201.

Macmahon, S., Peto, R., Cutler, J., *et al.* (1990). Blood-pressure, stroke, and coronary heart-disease .1. prolonged differences in blood-pressure - prospective observational studies corrected for the regression dilution bias. *Lancet* **335**, 765-774.

Manning, W., and Mullahy, J. (2001). Estimating log models: to transform or not to transform. *Journal of Health Economics* **20**, 461-494.

Manning, W. G., Basu, A., and Mullahy, J. (2005). Generalized modeling approaches to risk adjustment of skewed outcomes data. *Journal of Health Economics* **24**, 465-488.

Marshall, J. R., Hastrup, J. L., and Ross, J. S. (1999). Mismeasurement and the resonance of strong confounders: Correlated errors. *American Journal of Epidemiology* **150**, 88-96.

McCullagh, P., and Nelder, J. A. (1989). *Generalized linear models,* 2nd edition. London ; New York: Chapman and Hall.

Messer, K., and Natarajan, L. (2008). Maximum likelihood, multiple imputation and regression calibration for measurement error adjustment. *Statistics in Medicine* **27**, 6332-6350.

Michels, K. B., Bingham, S. A., Luben, R., Welch, A. A., and Day, N. E. (2004). The effect of correlated measurement error in multivariate models of diet. *American Journal of Epidemiology* **160**, 59-67.

Muff, S., Riebler, A., Held, L., Rue, H. and Saner, P (2015). Bayesian analysis of measurement error models using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **64**, 231–252.

Natarajan, L., Flatt, S. W., Sun, X. Y., *et al.* (2006). Validity and systematic error in measuring carotenoid consumption with dietary self-report instruments. *American Journal of Epidemiology* **163**, 770-778.

Natarajan, L., Pu, M. Y., Fan, J. J., *et al.* (2010). Measurement Error of Dietary Self-Report in Intervention Trials. *American Journal of Epidemiology* **172**, 819-827.

Neuhouser, M. L., Di, C., Tinker, L. F., *et al.* (2013). Physical Activity Assessment: Biomarkers and Self-Report of Activity-Related Energy Expenditure in the WHI. *American Journal of Epidemiology* **177**, 576-585.

Nusser, S. M., Beyler, N. K., Welk, G. J., Carriquiry, A. L., Fuller, W. A., and King, B. M. N. (2012). Modeling Errors in Physical Activity Recall Data. *Journal of Physical Activity & Health* **9**, S56-S67.

Olsen, M., and Schafer, J. (2001). A Two-Part Random-Effects Model for Semicontinuous Longitudinal Data. *Journal of the American Statistical Association* **96**, 730-745.

Pickett, K. E., Rathouz, P. J., Kasza, K., Wakschlag, L. S., and Wright, R. (2005). Self-reported smoking, cotinine levels, and patterns of smoking in pregnancy. *Paediatric and Perinatal Epidemiology* **19**, 368-376.

Plasqui, G., Bonomi, A. G., and Westerterp, K. R. (2013). Daily physical activity assessment with accelerometers: new insights and validation studies. *Obesity Reviews* **14**, 451-462.

Plasqui, G., and Westerterp, K. R. (2007). Physical activity assessment with accelerometers: An evaluation against doubly labeled water. *Obesity* **15**, 2371-2379.

Plummer, M. (2003). JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. In *Distributed Statistical Computing (DSC 2003)*, 1-10. Vienna, Austria.

Raymond, H. M., Montgomery, D.C., Vining, G.G. (2010). *Generalized Linear Models With Applications in Engineering and the Sciences*, Second edition. Hoboken, New Jersey: John Wiley & Sons,Inc.,.

Riboli, E., Hunt, K. J., Slimani, N., *et al.* (2002). European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutrition* **5**, 1113-1124.

Riboli, E., and Kaaks, R. (1997). The EPIC project: Rationale and study design. *International Journal of Epidemiology* **26**, S6-S14.

Richardson, S., and Gilks, W. R. (1993). Conditional-independence models for epidemiologic studies with covariate measurement error. *Statistics in Medicine* **12**, 1703-1722.

Rodrigues-Motta, M., Galvis Soto, D. M., Lachos, V. H., *et al.* (2015). A mixed-effect model for positive responses augmented by zeros. *Stat Med*.

Rosner, B., and Gore, R. (2001). Measurement error correction in nutritional epidemiology based on individual foods, with application to the relation of diet to breast cancer. *American Journal of Epidemiology* **154**, 827-835.

Rosner, B., Michels, K. B., Chen, Y. H., and Day, N. E. (2008). Measurement error correction for nutritional exposures with correlated measurement error: Use of the method of triads in a longitudinal setting. *Statistics in Medicine* **27**, 3466-3489.

Rosner, B., Spiegelman, D., and Willett, W. C. (1990). Correction of Logistic-Regression Relative Risk Estimates and Confidence-Intervals for Measurement Error - the Case of

Multiple Covariates Measured with Error. *American Journal of Epidemiology* **132**, 734-745.

Rosner, B., Willett, W. C., and Spiegelman, D. (1989). Correction of Logistic-Regression Relative Risk Estimates and Confidence-Intervals for Systematic within-Person Measurement Error. *Statistics in Medicine* **8**, 1051-1069.

Rothman, K. J., Greenland, S., and Lash, T. L. (2008). *Modern epidemiology*, 3rd edition. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins.

Royston, P., and Sauerbrei, W. (2003). Stability of multivariable fractional polynomial models with selection of variables and transformations: a bootstrap investigation. *Statistics in Medicine* **22**, 639-659.

Royston, P., and Sauerbrei, W. (2004). A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. *Statistics in Medicine* **23**, 2509-2525.

Royston, P., and Sauerbrei, W. (2005). Building multivariable regression models with continuous covariates in clinical epidemiology - With an emphasis on fractional polynomials. *Methods of Information in Medicine* **44**, 561-571.

Royston, P., and Sauerbrei, W. (2008). Multivariable model-building : a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Chichester, England ; Hoboken, NJ: John Wiley.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons,Inc.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*. Hoboken, N.J. ;: Wiley-Interscience.

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

Sainani, K. L. (2009). Linearity assessment in the Cox model. Linearity assessment in the Cox model. Stanford University.

Sasaki, J. E., John, D., and Freedson, P. S. (2011). Validation and comparison of ActiGraph activity monitors. *Journal of Science and Medicine in Sport* **14**, 411-416.

Schoeller, D. A., Leitch, C. A., and Brown, C. (1986). Doubly Labeled Water Method - Invivo Oxygen and Hydrogen Isotope Fractionation. *American Journal of Physiology* **251**, R1137-R1143.

Secker-Walker, R. H., Vacek, P. M., Flynn, B. S., and Mead, P. B. (1997). Exhaled carbon monoxide and urinary cotinine as measures of smoking in pregnancy. *Addictive Behaviors* **22**, 671-684.

Shibata, A., Paganinihill, A., Ross, R. K., Yu, M. C., and Henderson, B. E. (1992). Dietary Beta-Carotene, Cigarette-Smoking, and Lung-Cancer in Men. *Cancer Causes & Control* **3**, 207-214.

Slater, B., Enes, C. C., Lopez, R. V. M., Damasceno, N. R. T., and Voci, S. M. (2010). Validation of a food frequency questionnaire to assess the consumption of carotenoids, fruits and vegetables among adolescents: the method of triads. *Cadernos De Saude Publica* **26**, 2090-2100.

Slimani, N., Kaaks, R., Ferrari, P., *et al.* (2002). European Prospective Investigation into Cancer and Nutrition (EPIC) calibration study: rationale, design and population characteristics. *Public Health Nutrition* **5**, 1125-1145.

Slimani, N., Valsta, L., and Grp, E. (2002). Perspectives of using the EPIC-SOFT programme in the context of pan-European nutritional monitoring surveys: methodological and practical implications. *European Journal of Clinical Nutrition* **56**, S63-S74.

SmithWarner, S. A., Elmer, P. J., Fosdick, L., Tharp, T. M., and Randall, B. (1997). Reliability and comparability of three dietary assessment methods for estimating fruit and vegetable intakes. *Epidemiology* **8**, 196-201.

Stefanski, L. A., and Cook, J. R. (1995). Simulation-extrapolation: The measurement error jackknife. *Journal of the American Statistical Association* **90**, 1247-1256.

Steyerberg, E. W. (2009). Clinical prediction models : a practical approach to development, validation, and updating. New York, NY: Springer.

Stram, D. O., Huberman, M., and Wu, A. H. (2002). 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**, 622-628.

Subar, A. F., Kipnis, V., Troiano, R. P., *et al.* (2002). Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: The observing protein and energy nutrition study. *Faseb Journal* **16**, A27-A27.

Subar, A. F., Kipnis, V., Troiano, R. P., *et al.* (2003). Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: The OPEN Study. *American Journal of Epidemiology* **158**, 1-13.

Swartz, A. M., Strath, S. J., Bassett, D. R., O'Brien, W. L., King, G. A., and Ainsworth, B. E. (2000). Estimation of energy expenditure using CSA accelerometers at hip and wrist sites. *Medicine and science in sports and exercise* **32**, S450-S456.

Thomas, D., Stram, D., and Dwyer, J. (1993). Exposure measurement error - influence on exposure-disease relationships and methods of correction. *Annual Review of Public Health* **14**, 69-93.

Thoresen, M. (2006). Correction for measurement error in multiple logistic regression : A simulation study. *Journal of Statistical Computation and Simulation,* **76**, 475-487.

Thoresen, M., and Laake, P. (2000). A simulation study of measurement error correction methods in logistic regression. *Biometrics* **56**, 868-872.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B-Methodological* **58**, 267-288.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in Medicine* **16**, 385-395.

Tooze, J., Midthune, D., Dodd, K., *et al.* (2006). A new method for estimating the usual intake of episodically consumed foods with application to their distribution. *Journal of American Diet Association*, 1575-1587.

Tooze, J. A., Troiano, R. P., Carroll, R. J., Moshfegh, A. J., and Freedman, L. S. (2013). A Measurement Error Model for Physical Activity Level as Measured by a Questionnaire With Application to the 19992006 NHANES Questionnaire. *American Journal of Epidemiology* **177**, 1199-1208.

Trijsburg, L., de Vries, J. H., Boshuizen, H. C., *et al.* (2015). Comparison of duplicate portion and 24 h recall as reference methods for validating a FFQ using urinary markers as the estimate of true intake. *Br J Nutr*, 1-9.

Van Remoortel, H., Giavedoni, S., Raste, Y., *et al.* (2012). Validity of activity monitors in health and chronic disease: a systematic review. *Int J Behav Nutr Phys Act* **9**, 84.

van Rossum, C., Fransen, HP., Verkaik-Kloosterman, J., Buurma-Rethans, EJM., Ocké, MC. (2011). Dutch National Food Consumption Survey 2007-2010: : Diet of children and adults aged 7 to 69 years. Bilthoven: RIVM.

Weir, J. B. D. (1949). New Methods for Calculating Metabolic Rate with Special Reference to Protein Metabolism. *Journal of Physiology-London* **109**, 1-9.

Weiss, J. (2006). Statistical Analysis. In *Ecology 145—Statistical Analysis*. University of North Carolina, Chapel Hill: University of North Carolina.

WHO (2015). Physical Activity. WHO.

Willet, W. (1998). *Nutritional Epidemiology*. New York: Oxford University Press.

Wong, M. Y., Day, N. E., and Wareham, N. J. (1999). Measurement error in epidemiology: The design of validation studies - II: Bivariate situation. *Statistics in Medicine* **18**, 2831-2845.

Wood, A. M., White, I. R., Thompson, S. G., *et al.* (2009). Correcting for multivariate measurement error by regression calibration in meta-analyses of epidemiological studies. *Statistics in Medicine* **28**, 1067-1092.

Wood, S. N. (2012). Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation

Woodward, M., Moohan, M., and Tunstall-Pedoe, H. (1999). 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**, 285-295.

Zhang, S., Krebs-Smith, S. M., Midthune, D., *et al.* (2011). Fitting a bivariate measurement error model for episodically consumed dietary components. *Int J Biostat* **7**, 1.

Zhang, S., Midthune, D., Guenther, P. M., *et al.* (2011). A New Multivariate Measurement Error Model with Zero-Inflated Dietary Data, and Its Application to Dietary Assessment. *Ann Appl Stat* **5**, 1456-1487.

Zhao, S. S., and Prentice, R. L. (2014). Covariate Measurement Error Correction Methods in Mediation Analysis with Failure Time Data. *Biometrics* **70**, 835-844.

Summary

Measurement error in exposure variables is an important factor in epidemiological studies. Epidemiologic studies relate exposures, such as dietary intakes and level of physical activity, to health outcomes. Such studies, however, usually pay limited attention to the quantitative effects of exposure measurement error on the estimated association between an exposure and a health outcome. Measurement error in the exposure usually leads to a biased estimate of the parameter that quantifies the exposure-outcome association.

Methods to adjust for such measurement error require a validation study. To adjust for measurement error in exposures that are validated using zero-inflated reference measurements with the existing methods, a validation study with multiple replicates of the reference measurement is usually required; for instance, episodically consumed foods measured by short-term instruments such as 24-hour recall. Validation studies with multiple replicates are quite costly. Hence, in some cases either a single–replicate or no validation study is conducted besides the main study.

In this thesis, we adapted regression calibration to adjust for exposure measurement error for a single-replicate validation study with zero-inflated reference measurements and assessed the adequacy of the adapted method in a simulation study. For the case where there is no internal validation study, we showed how to combine external validation data on validity for self-report instruments with the observed questionnaire data to adjust for the bias in the associations caused by measurement error in correlated exposures. In the last part, we assessed the effect of measurement error in physical activity as measured by accelerometer in a recently concluded validation study.

In **chapter 2**, a two-part regression calibration model was adapted to the case of a single-replicate validation study with zero-inflated reference measurements. The chapter describes how the excess zero values in the response were handled using a two-part modelling approach, where the first part models the probability of a non-zero response and the second part conditionally models the response given that it is non-zero. Further described is how the optimal parametric forms of the skewed continuous covariates in each part of the calibration model were explored using generalized additive modelling and empirical logit approaches, and how the covariates were selected into the calibration model. The performance of the proposed two-part model was compared with less specified calibration models for episodically consumed dietary intakes measured with error in a real dataset. The performance of the proposed method was influenced more by the functional forms of continuous covariates than the assumed distribution for the response in the calibration model. It is concluded from this chapter that when only a single-replicate validation study with zero-inflated reference measurements is available, a correctly specified regression calibration can be used to adjust for the bias in the exposure-outcome associations.

In **chapter 3**, a simulation study was conducted to assess the performance of the model proposed in chapter 2. The model was assessed with respect to the magnitude of the correlation between the probability of a non-zero response and the actual non-zero response values (cross-part correlation), the number and form of covariates in the calibration model, the percentage of zeroes in the calibration response, and the magnitude of the measurement error in the exposure. The model performance was minimally influenced by the cross-part correlation, and was more sensitive to the form of continuous covariate than the assumed distribution for the response. Reducing the number of covariates in the model seemed beneficial, but was not critical in large-sample studies.

In **chapter 4**, a multivariate method was proposed to adjust for the bias in the exposure-outcome association caused by measurement error in the exposure and confounder variables, in the absence of an internal validation study. In the proposed method, Bayesian Markov chain Monte Carlo technique was used to combine external validation data for self-report questionnaires with the observed questionnaire data to adjust for the bias in the association. The method was compared with a method that

ignores measurement error in the confounder. Further, a sensitivity analysis was performed to get insight into the influence of assumptions on the measurement error structure, mainly with respect to the magnitude of error and correlation between the errors. The proposed method was illustrated with a real dataset. Via sensitivity analysis, it was shown that due to measurement error in the confounder, the exposure-outcome associations can still be biased, even when the exposure is measured without error. Additionally, if the confounder is strongly linked with the outcome, measurement error in the confounder can be more influential than measurement error in the exposure in causing the bias in the exposure-outcome association; moreover, the bias can be in any direction. We concluded that when there is no internal validation study, carefully elicited external validation data that is transportable to the main study can be used to adjust for the bias in the exposure-outcome associations. The proposed method is also useful in conducting sensitivity analyses on the effect of measurement errors in the exposure and confounder variables and error correlations on the observed exposure-outcome association.

In **chapter 5**, a triaxial accelerometer (GT3X) was validated against doubly labelled water for assessing physical activity in free-living individuals in a recently concluded validation study. We applied a measurement error model and quantified measurement error with: (i) the bias in the mean level of physical activity, (ii) the correlation coefficient between measured and true level of physical activity to quantify loss of statistical power to detect associations, and (iii) the attenuation factor to quantify the bias in the associations between physical activity and health outcomes. We showed that when accelerometers, such as GT3X, are used in a population similar to that presented in chapter 5 to monitor the level of physical activity in free-living individuals, the mean level of physical activity would be underestimated, the associations between physical activity and health outcomes would be biased, typically towards the null, and there would be loss of statistical power to detect associations. From this chapter we made two important remarks. First, if physical activity measurements from GT3X accelerometer are used to study the association between physical activity and a health outcome, the derived conclusions may be misleading. Second, when a biased "reference" instrument, such as the GT3X accelerometer, is used to validate a main-study instrument, such as a physical activity questionnaire, the true effect of measurement error in the latter would be misrepresented.

In **chapter 6**, the main findings, limitations and their implications are discussed, followed by suggestions for improvement and potential areas for future research. The following are the main concluding remarks. First, to adjust for the bias in the association adequately, the proposed two-part calibration should be specified correctly. Second, the performance of the proposed calibration model is influenced more by the assumption made on the form of the continuous covariates than the form of the response distribution. Third, in the absence of an internal validation study, carefully extracted validation data that is transferrable to the main study can be used to adjust for the bias in the associations. Lastly, when "reference" instruments are themselves marred by substantial bias, the effect of measurement error in an instrument being validated can be seriously underestimated.

Acknowledgements

To my supervisors: Hendriek Boshuizen, Fred van Eeuwijk and Hilko van der Voet, thank you very much for your invaluable guidance and support that led to the successful completion of this thesis. To Hendriek and Hilko, I would not be writing this acknowledgement had it not been for those fruitful weekly meetings and discussions on my thesis work. To Fred, I deeply express my gratitude for your insightful comments and for introducing me to the International Biometric Society. My appreciations go to Paul Goedhart for taking your time to assist me with the simulation study design and for those insightful suggestions. Many thanks to Pieter van' t Veer for your guidance, advices and insightful suggestions on my thesis work. Thank you very much Pieter for your time and invaluable contributions during our bimonthly progress meetings.

I consider myself blessed to have done my PhD at Biometris Department. Thanks to everyone at Biometris. You made me feel at home. I am very grateful to Jaap Molenaar for your words of encouragement and for the numerous recommendation letters that you wrote for me despite your busy schedule. To Cajo ter Braak, I highly appreciate your advices and insightful suggestions on my work during the PhD days. Thank you Sabine Schnabel for encouraging me to join and participate in the International Workshop on Statistical Modelling Society. Many thanks to Hans Stigter, Joao, Gerrit, Lia, Elly, Ron, Maikel, Maarten de Gee, Marcos, Marco, Bas, Bakker, Waldo, Joost, Johannes, Onno, Willem, Chaozhi, Paul Eilers, Cristian, Paul Keizer, Martin Boer, Eric Boer, Saskia and all other members of Biometris for making me feel at home away from home. I highly appreciate every conversation that we had during my stay at Biometris.

My stay at Biometris would not have been easy without the support of two wonderful secretaries: Dinie Verbeek and Hanneke van Ommeren-Keuken. Thank you very much for those Dutch translations, travel arrangements and for every way that you supported me.

Many thanks to my fellow PhD colleagues: Rianne Jacobs, Daniela, Vincent, Martin, Guus, Nadia, Rianne van Binsbergen, Jeroen, Cassandra, Amber, Sanne, Julio, Tryntsje, Maggie and Nurudeen. You made my stay at Biometris special and I would not have asked for better colleagues than you guys. Special thanks to Rianne Jacobs for those insightful comments and for helping me with thesis formatting. I express my deep gratitude to Rianne and Daniela for helping me with the defence preparation and for taking your time to proof-read my thesis.

I had a wonderful time at RIVM. I am very grateful to Hendriek Boshuizen, Jose Ferreira, Albert Wong, Jan van de Kassteele, Jack Schijven, Maarten Schipper and everyone at RIVM. To Jose Ferreira, I learnt a lot from you and many thanks for those inspirational conversations and for the books. You are such a great guy!

My appreciations go to the DuPLO study group for permitting me to use the physical activity data. Laura Trijsburg, thank you for your support and in helping me analyse the accelerometry data. I am very grateful to the EPIC collaborators, especially Pietro Ferrari for your comments on my manuscripts.

Thank you very much Kristel Luijtens for your continual guidance and for your willingness to be my referee.

To my Kenyan community in Wageningen, thank you very much guys for those *nyama choma* moments and parties. You guys made me feel very much at home. Many thanks to Yussuf, Jacky, Faith, Okayo, Hellen, Winnie, Steve, Martin, Maati, Irine, Collins, Sophy and all WUR alumni. To JKUAT colleagues, I am very grateful for your support and friendship.

Many thanks to my measurement error group members: Hendriek, Laura and Olga for those insightful discussions. I am grateful to my running team members: Guus, Faith, Winnie, Stella, Claire, Julio, Chaozhi, Yussuf and Okita for your support. Special thanks to all members of AGP church for my spiritual nourishment and growth. To house fellowship members Elton, Ednah, Gerbert, Janet, Claire, Ismael and others, thank you very much for those bible study sessions. I am indebted to the family of Capt. Kopondo for your immeasurable support over the years. Thank you very much Capt. Kopondo for that memorable flight from Amsterdam to Nairobi.

I am very grateful to my special friend. Thank you very much for all the sacrifices and for supporting me throughout my PhD journey. It was not easy, but we made it. Special thanks go to my family for the prayers and support throughout this journey. Many thanks to my sisters Florence, Ann, Debora and Gaudenciah, bro Michael (RIP), Phillis, the little girls: Kristy and Karen, Peter Jamwa, John Jinnah and other family members and relatives for your support.

It would not be possible without the support of the most influential woman in my life—mama Christine. Thank you very much mum for your prayers, words of encouragement, support and all kinds of sacrifices that you went through just to make this dream a reality. I admire your strength mum and I am more than blessed to be your son. May you be richly blessed.

About the Author

George Onyango Agogo was born on 15th February, 1984 in Bondo, Kenya. George holds a Bachelor degree in Mathematics and Computer Science (First Class Honours) from Jomo Kenyatta University of Agriculture and Technology (2008), Kenya. He then joined Safaricom (a telecommunication company in Kenya) till September 2009. George applied for VLIR (*Vlaamse Interuniversitaire Raad*) scholarship to study in Belgium, which he was awarded in 2009. In October, 2009, George started his MSc in Biostatistics at Hasselt University, Belgium and graduated with Distinction in September, 2011. The title of his MSc thesis was "Certolizumab Pegol (Cimzia) long term treatment outcome predicted by short term results in patients with Rheumatoid Arthritis" under the supervision of Dr. Cristina Sotto of Hasselt University and Dr. Kristel Luijtens of UCB Pharma, Brussels, Belgium. George also held teaching assistant position at JKUAT, where he assisted Bachelor students with tutorials in Mathematics and Statistics during his summer vacation. In September, 2011, George was awarded a PhD position by Wageningen University and Research Centre in collaboration with RIVM, Netherlands. He started his PhD on 1 November, 2011 under the supervision of Prof. dr. Hendriek Boshuizen, Prof. dr. Fred van Eeuwijk and Dr. Hilko van der Voet. During his PhD, George presented his scientific work in several international conferences. He won three paper competition awards: one awarded by World Bank Fund and the other two awarded by Statistics South Africa (Stats SA). Besides his PhD work, George served as the chairman of Kenyans in Wageningen student organization from 2012 to 2013. Throughout his PhD period, he assisted many students (Masters and PhDs) with statistical analyses. His long-term plan is to pursue a career in a health-related sector.

List of publications

Agogo, G.O., van der Voet, H., van 't Veer, P., Ferrari, P., et al. (2014). Use of Two-Part Regression Calibration Model to Correct for Measurement Error in Episodically Consumed Foods in a Single-Replicate Study Design: EPIC Case Study. PLoS ONE 9(11): e113160. doi:10.1371/journal.pone.0113160.

Agogo, G.O., van der Voet, H., van Eeuwijk, F.A. and Boshuizen, H.C.(2015). Evaluation of two-part regression calibration to adjust for dietary exposure measurement error in the Cox proportional hazards model: a simulation study. Biometrical journal (In press).

Agogo, G.O., van der Voet, H., Trijsburg, L., van Eeuwijk, F.A., et al.(2015). Measurement error modelling for accelerometer activity data using Bayesian integrated nested Laplace approximation. In *Proceedings of the 30th International Workshop on Statistical Modelling* 2(pp. 3-6). Linz, Austria: Johannes Kepler University.

Agogo, G.O., van der Voet, H., van Eeuwijk, F.A.,van ’t Veer, P., et al. Multivariate bias adjustment method to deal with measurement error in dietary intake using external validation data (submitted).

Agogo, G.O., van der Voet, H., Hulshof, P.J.M., van 't Veer, P.,et al.(2015).Quantification of Measurement Error in Accelerometer activity Data (to be submitted).

Conference abstracts

Agogo, G.O., van der Voet, H., van 't Veer, P., Ferrari, P., et al. Use of Two-Part Regression Calibration Models for Intake Data measured with many zeros and error: EPIC Case study-IARC-WHO, LYON, France:-Statistics Meeting, November, 13-15, 2012 (Oral presentation)

Agogo, G.O., van der Voet, H., van 't Veer, P., van Eeuwijk, F., and Boshuizen H.C. A two-Part Regression Calibration Models to correct for measurement error: A simulation study, University of St Andrews, Scotland, United Kingdom:-4th CNC IBS-BIR Netherlands Region Conference, July, 3-5, 2013 (Poster presentation)

Agogo, G.O. A flexible zero-augmented generalized gamma random effects regression calibration model to correct for measurement error in the exposure variable: Florence, Italy:-27th IBC, July 6-12, 2014 (Oral presentation)

Agogo, G.O. A flexible two-part generalized gamma regression calibration model to correct for measurement error in the exposure variable: Pretoria, South Africa:- 4th ISIbalo IYASC, July 31-August 2, 2014 (Oral presentation)

Agogo, G.O., van der Voet, H., Hulshof, P.J.M., van 't Veer, P.,et al. Measurement error models for accelerometer data using MCMC, Integrated Nested Laplace Approximations and Maximum Likelihood Estimation methods: Nijmegen, Netherlands:- 5th IBS-BIR and Netherlands Region, April 20-22,2015 (Poster presentation)

Agogo, G.O., van der Voet, H., Hulshof, P.J.M., van 't Veer, P.,et al. A measurement error model for accelerometer activity data using Integrated Nested Laplace Approximations: Linz, Austria:- 30th International Workshop on Statistical Modelling, July 6-10,2015 (Poster presentation)

Agogo, G.O. A flexible Zero-Augmented Generalized Gamma Mixed Effects Regression Calibration model to Correct for Measurement Error in Episodically Consumed Food: Rio de Janeiro, Brazil:- 60th ISI World Statistics Congress, July 26-31, 2015 (Oral presentation)

Overview of completed training activities

Discipline specific activities
Courses

- Exposure Assessment in Nutrition Research, 2012, Wageningen, Netherlands
- Public Health Research in Practice: How to develop effective interventions in public health practice?, 2013, Wageningen, Netherlands
- Nutritional and lifestyle epidemiology, 2013, Wageningen, Netherlands
- Zero-inflated models with GLMM in R, 2014, Wageningen, Netherlands

Conferences and Workshops

- EPIC Statistics meeting, 2012, Lyon, France
- 4th International Biometric Society (IBS) Channel meeting, 2013, St Andrews, UK
- 27th IBS conference, 2014, Florence, Italy
- 4th ISIbalo Young African Statistician Conference, 2014, Pretoria, South Africa
- 5th IBS Channel meeting, 2015, Nijmegen, Netherlands
- 30th International Workshop on Statistical Modelling, 2015, Linz, Austria
- 60th ISI World Statistics Congress, 2015, Rio de Janeiro, Brazil

General courses

- VLAG PhD Week courses, 2012, Baarlo, Netherlands
- Scientific Publishing, 2013, Wageningen, Netherlands
- Techniques of writing and presenting a scientific paper, 2014, Wageningen, Netherlands
- Scientific Integrity, 2012, Wageningen, Netherlands
- Biometris PhD workshops, 2013, 2014, Wageningen
- Scientific writing, 2014, Wageningen, Netherlands
- Introduction to LateX, 2012, Wageningen, Netherlands

Optional courses and activities

- SB&E colloquium, 2013-2014, Wageningen, Netherlands
- Measurement error modelling (webinar series), 2011, RIVM, Netherlands
- Measurement error group, 2012-2014, Wageningen, Netherlands
- Concepts and methods in epidemiology (Rothman lunches), 2012/2013, Wageningen, Netherlands
- Exposure Assessment in Nutrition and Health Research (MSc course), 2012, Wageningen, Netherlands
- Preparation of Research Proposal, 2012, Wageningen, Netherlands
- Design and Interpretation of epidemiological studies (MSc Course), 2013, Wageningen, Netherlands
- Assisting with practical for Basic statistics (MSc course), 2013, Wageningen, Netherlands

Funding

The research in this thesis was financially supported by Wageningen University and National Institute of Public Health and the Environment (RIVM) in Bilthoven.

The cover of this thesis was designed by George O. Agogo and GVO drukkers & vormgevers.