Received: 12 February 2019 | Revised: 9 September 2019 | Accepted: 14 September 2019
DOI: 10.1002/bimj.201900066

Biometrical Journal

RESEARCH PAPER

SIMEX for correction of dietary exposure effects with Box-Cox transformed data

Timm Intemann[1,2] | Kirsten Mehlig[3] | Stefaan De Henauw[4] | Alfonso Siani[5] | Tassos Constantinou[6] | Luis A. Moreno[7] | Dénes Molnár[8] | Toomas Veidebaum[9] | Iris Pigeot[1,2] | on behalf of the I.Family consortium

[1]Biometry and Data Management, Leibniz Institute for Prevention Research and Epidemiology–BIPS, Bremen, Germany

[2]Institute of Statistics, Faculty of Mathematics and Computer Science, University of Bremen, Bremen, Germany

[3]Department of Public Health and Community Medicine, Institute of Medicine, Sahlgrenska Academy at University of Gothenburg, Gothenburg, Sweden

[4]Department of Public Health, Ghent University, Ghent, Belgium

[5]Institute of Food Sciences, National Research Council, Avellino, Italy

[6]Research and Education Institute of Child Health, Strovolos, Cyprus

[7]GENUD (Growth, Exercise, Nutrition and Development) Research Group, Instituto Agroalimentario de Aragón (IA2), Instituto de Investigación Sanitaria Aragón (IIS Aragón), Centro de Investigación Biomédica en Red Fisiopatologa de la Obesidad y Nutrición (CIBERObn), University of Zaragoza, Zaragoza, Spain

[8]Department of Pediatrics, Medical School, University of Pécs, Pécs, Hungary

[9]Department of Chronic Diseases, National Institute for Health Development, Tallinn, Estonia

Correspondence
Timm Intemann, Biometry and Data Management, Leibniz Institute for Prevention Research and Epidemiology–BIPS, Achterstr. 30, 28359 Bremen, Germany.
Email: intemann@leibniz-bips.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: 391977161; European Commission Seventh Framework Programme, Grant/Award Number: 266044

Abstract

Modelling dietary data, and especially 24-hr dietary recall (24HDR) data, is a challenge. Ignoring the inherent measurement error (ME) leads to biased effect estimates when the association between an exposure and an outcome is investigated. We propose an adapted simulation extrapolation (SIMEX) algorithm for modelling dietary exposures. For this purpose, we exploit the ME model of the NCI method where we assume the assumption of normally distributed errors of the reported intake on the Box-Cox transformed scale and of unbiased recalls on the original scale. According to the SIMEX algorithm, remeasurements of the observed data with additional ME are generated in order to estimate the association between the level of ME and the resulting effect estimate. Subsequently, this association is extrapolated to the case of zero ME to obtain the corrected estimate. We show that the proposed method fulfils the key property of the SIMEX approach, that is, that the MSE of the generated data will converge to zero if the ME variance converges to zero. Furthermore, the method is applied to real 24HDR data of the I.Family study to correct the effects of salt and alcohol intake on blood pressure. In a simulation study, the method is compared with the NCI method resulting in effect estimates with either smaller MSE or smaller bias in certain situations. In addition, we found our method to be more informative and easier to implement. Therefore, we conclude that the proposed method is useful to promote the dissemination of ME correction methods in nutritional epidemiology.

KEYWORDS

24-hr dietary recall, measurement error correction method, NCI method, non-linear mixed model, salt intake

1 | INTRODUCTION

Measurement error (ME) can lead to seriously wrong conclusions about associations between exposures and health outcomes (Carroll, Ruppert, Stefanski, & Crainiceanu, 2006). Correction methods can reduce the negative consequences of ME. These methods play a big role in modelling exposures based on data from dietary assessment tools as from the 24-hr dietary recall (24HDR) (Souverein et al., 2011). The 24HDR is used to repeatedly record the participant's dietary intake of entire days based on self-reports delivered on the following day. Since the self-reports are both error-prone and assessed on a daily basis, the data are strongly affected by intra-individual variation and do not reflect the usual intake nutritional epidemiologist are primary interested in (Boeing & Margetts, 2014). Furthermore, the intake distributions are usually positively skewed and sometimes zero-inflated.

The most commonly used method for modelling exposures based on 24HDR data that accounts for their special characteristics is the NCI method (Tooze et al., 2006; Kipnis et al., 2009). This method follows the regression calibration approach which consists of two parts. First, following a ME model for 24HDR data, the estimated conditional expectation of the usual intake given the observed 24HDR and other covariates is calculated. The ME model accounts for the skewness of the intake distributions by using the Box-Cox transformation. Second, in a health model, the estimated usual intake is used instead of the unknown true usual intake. The assumed ME model of the NCI method describes the association between the true and measured intake and the health model describes the association between usual intake and health outcome. Therefore, the NCI method provides nearly unbiased estimates for the association between dietary intake and health outcome (Kipnis et al., 2009) and has been used in various studies (Börnhorst et al., 2014; Liese et al., 2015; Hebestreit et al., 2017; Intemann et al., 2018).

However, a recent study by Shaw et al. (2018) about the usage of ME correction methods shows that studies with inadequate correction for ME remain common in nutritional epidemiology. Following the Measurement Error and Misclassification Topic Group of the STRATOS Initiative, it is important to raise awareness for ME problems and to further promote correction methods and their use (Freedman & Kipnis, 2018).

The simulation extrapolation (SIMEX) method (Cook & Stefanski, 1994; Stefanski & Cook, 1995), a promising, clear, and easy to implement correction method, has not been used for modelling 24HDR data, although it is, aside from regression calibration, one of the most prominent approaches. Basically, this method can be used in any situation where the underlying error model can be simulated by Monte Carlo methods. The idea of SIMEX is to add well-defined error terms to the observed variable, determine its association with a health outcome, and extrapolate the resulting effect estimates back to zero ME. For this purpose, remeasurements of the original data with varying level of additional error are generated. Based on the generated data, effect estimates are calculated and the functional association between these estimates and the level of ME is estimated. For some common error models, SIMEX is implemented in statistical software packages, for example, in the `R` package `simex` (Lederer & Küchenhoff, 2006).

Therefore, the aim of this paper is to adapt the SIMEX algorithm for a model of 24HDR data, to provide a theoretical justification for SIMEX in this situation, and to investigate the algorithm in an application as well as in a simulation study. The paper is structured as follows. Section 2 describes the error and health model throughout. In Sections 3, 4, and 5, the SIMEX algorithm is introduced in the context of 24HDR data, its properties are discussed and its extension for episodically consumed foods is given. The proofs of the properties can be found in Appendices A.1 and A.2. The focus of these proofs is on the mean squared error (MSE) for the generated SIMEX data in case that the ME converges to zero. In Section 6, the SIMEX algorithm is applied to real 24HDR data of the I.Family study (Ahrens et al., 2017) to investigate the association between salt and alcohol intake and blood pressure. These studies serve as template for the simulation study in Section 7 where the SIMEX algorithm is compared with the NCI method. Finally, the results are discussed.

2 | MEASUREMENT ERROR AND HEALTH MODEL

The ME and health model for 24HDR data used in this paper were developed by Dodd et al. (2006), Tooze et al. (2006), and Kipnis et al. (2009). The reported intake of individual i on day j is denoted as R_{ij} and the true individual usual intake as T_i, $i = 1, \dots, I, j = 1, \dots, J_i$. We assume that R_{ij} is unbiased for T_i on the original scale following the convention in nutritional epidemiology (Dodd et al., 2006). Furthermore, we assume that (i) there exists a Box-Cox transformation $g_\lambda(v) = (v^\lambda - 1)/\lambda$ if $\lambda > 0$ and $g_\lambda(v) = \log(v)$ if $\lambda = 0$ such that

$$g_\lambda(R_{ij}) = \mathrm{E}(g_\lambda(R_{ij})) + \varepsilon_{ij} \tag{1}$$

with independent random variables $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, and (ii) the regression model $\mathrm{E}(g_\lambda(R_{ij})) = \beta_0 + \boldsymbol{\beta}\mathbf{X}_i + u_i$ holds, where $u_i \sim \mathcal{N}(0, \sigma_u^2)$ are independent random variables, β_0 and $\boldsymbol{\beta}$ are parameters, and $\mathbf{X}_i$ are vectors of error free covariates. Combing both gives the non-linear mixed effects error model

$$g_\lambda(R_{ij}) = \beta_0 + \boldsymbol{\beta}\mathbf{X}_i + u_i + \varepsilon_{ij}. \tag{2}$$

The random variables u_i and ε_{ij} are assumed to be mutually independent. The former reflects the inter-individual and the latter the intra-individual variation. It is important to note that $g_\lambda(R_{ij})$ is biased for $g_\lambda(T_i)$ since it is assumed that R_{ij} is *unbiased* for T_i on the original scale.

For modelling the association between the usual intake T_i and a health outcome H_i the so-called health model is defined as the following regression model

$$\mathrm{E}(H_i | T_i, \mathbf{X}'_i) = \alpha_0 + \alpha_T T_i + \boldsymbol{\alpha}_X \mathbf{X}'_i \tag{3}$$

with the parameters α_0, α_T, and $\boldsymbol{\alpha}_X$ and the error free covariates $\mathbf{X}'_i$ which are all included in $\mathbf{X}_i$. If the error model is ignored and the individual mean $\bar{R}_{i\bullet}$ is used instead of T_i in the naïve health model

$$\mathrm{E}(H_i | \bar{R}_{i\bullet}, \mathbf{X}'_i) = \alpha_0 + \alpha_R \bar{R}_{i\bullet} + \boldsymbol{\alpha}_X \mathbf{X}'_i \tag{4}$$

the least-square estimator of α_R (the naïve estimator) will be biased for α_T. To reduce this bias, an adapted SIMEX algorithm is proposed in the next section.

3 | ADAPTED SIMEX ALGORITHM

The classical SIMEX algorithm is based on the assumption of normally distributed error on the original scale. Therefore, an adaption is necessary to account for the transformation g_λ in error model (2). The proposed adapted algorithm consists of the following steps:

(i) If λ and σ_ε^2 are unknown, the parameters of the non-linear mixed model (2) will be estimated using a maximum likelihood (ML)-like and a restricted ML approach (for details see Appendix A.3).

(ii) Following Carroll et al. (2006) in the simulation step, the equation

$$R_{ij}^{(l)}(\zeta) = g_\lambda^{-1}\Big(g_\lambda(R_{ij}) + Z_{ij}^{(l)}(\zeta)\Big) \tag{5}$$

is used to generate remeasurements of R_{ij} in the l-th data set, $l = 1, \dots, L$. The function $g_\lambda^{-1}(v) = (\lambda v + 1)^{1/\lambda}$ if $\lambda > 0$ and $g_\lambda^{-1}(v) = \exp(v)$ if $\lambda = 0$ is the inverse function of g_λ, the pseudo-random variable $Z_{ij}^{(l)}(\zeta)$ follows a normal distribution with $\mathcal{N}(\mu_i(\zeta), \zeta\sigma_\varepsilon^2)$ and $\zeta \in \mathcal{Z}$. For the choice of values for L and $\mathcal{Z}$ see, for example, the study in Section 4. For the calculation of the corrective expected value $\mu_i(\zeta)$, we refer to Section 4. It guarantees for each individual that $R_{ij}^{(l)}(\zeta)$ is (approximately) unbiased for R_{ij} on the original scale.

Using the Box-Cox transformation with $\lambda > 0$ Equation (5) leads to

$$R_{ij}^{(l)}(\zeta) = \Big(R_{ij}^\lambda + \lambda Z_{ij}^{(l)}(\zeta)\Big)^{1/\lambda}. \tag{6}$$

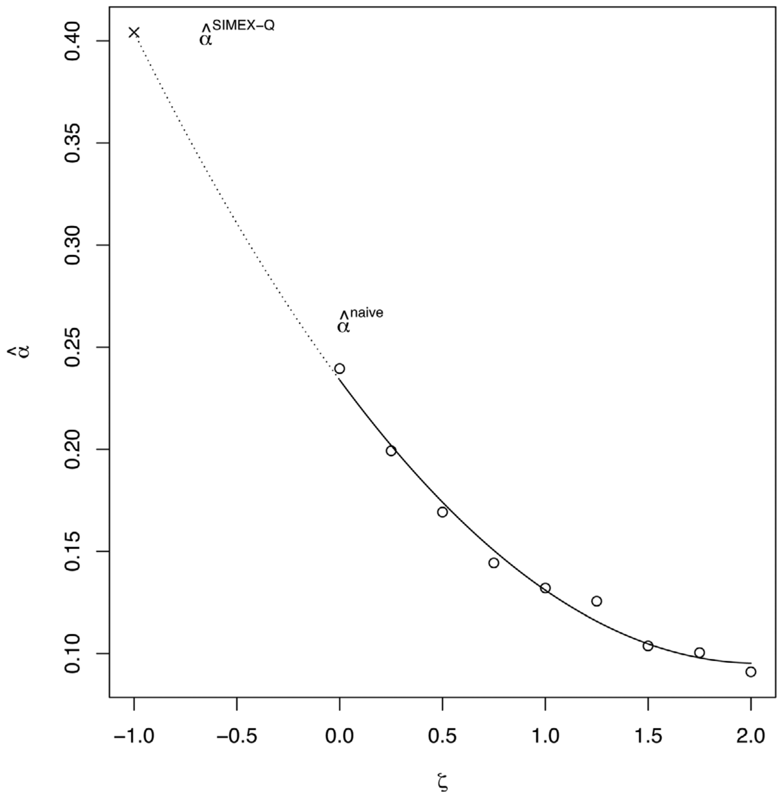

FIGURE 1 The typical SIMEX plot shows the naïve effect estimate (at $\zeta = 0$), the mean effect estimates for $\zeta \in \mathcal{Z}$ and the SIMEX-Q corrected estimate $\hat{\alpha}^{\text{SIMEX}-Q}$ (at $\zeta = -1$) which is based on the plotted quadratic extrapolation function

(iii) The naïve health model (4) where $\bar{R}_{i\bullet}$ is replaced by $\bar{R}_{i\bullet}^{(l)}(\zeta)$ is fitted to the $\#\mathcal{Z} \times L$ generated data sets. The corresponding effect estimates are denoted as $\hat{\alpha}_R^{(l)}(\zeta)$.

(iv) This and the following steps are conducted according to the classical SIMEX algorithm. First, for each $\zeta \in \mathcal{Z}$, the arithmetic mean of the estimates $\hat{\alpha}_R^{(l)}(\zeta)$ of all generated data sets is calculated. It is denoted by $\hat{\alpha}_R(\zeta)$.

(v) Then, the $\hat{\alpha}_R(\zeta)$ are plotted against $\zeta \in \mathcal{Z}$. The estimate for α_R based on the naïve health model, $\hat{\alpha}_R$, is denoted as $\hat{\alpha}_R(0)$ and is plotted against $\zeta = 0$. Using a regression model, the association between $\zeta \in \mathcal{Z} \cup \{0\}$ as independent variable and the corresponding $\hat{\alpha}_R(\zeta)$ as dependent variable is estimated. For this purpose, the polynomial function $\mathcal{G}_k(\zeta) = \eta_1 + \eta_2\zeta^1 + \cdots + \eta_{k+1}\zeta^k$ or the rational linear function $\mathcal{G}_{RL}(\zeta) = \eta_1 + \eta_2/(\eta_3 + \zeta)$ are used. For $k = 2$ the algorithm is called SIMEX-Q, for $k = 3$ SIMEX-C, for $k = 4$ SIMEX-Q4, and for $k = 5$ SIMEX-Q5. If $\mathcal{G}_{RL}$ is used, the algorithm will be called SIMEX-RL, which has the following advantageous theoretical property under the classical error model. In a multiple linear regression model, the effect estimator depends on the error-prone variable and has the form of $\mathcal{G}_{RL}(\zeta)$ (for details see Carroll et al. 2006). However, the rational linear function has two drawbacks, (i) the success of the model fit depends on the choice of start parameters for η_1, η_2, and η_3 and (ii) $\eta_3 \in [0, 1]$ leads to a singularity for $\zeta \in [-1, 0]$ (Carroll et al., 2006).

(vi) The extrapolation step is the last step. The argument $\zeta = -1$ is inserted in the estimated function $\hat{\mathcal{G}}(\zeta)$. The result $\hat{\mathcal{G}}(-1) = \hat{\alpha}_R^{\text{SIMEX}}$ is denoted as SIMEX estimate. The justification for the extrapolation is given in the following section. If corrected effect estimates for the intercept and remaining covariates $\hat{\alpha}_0^{\text{SIMEX}}$ and $\hat{\boldsymbol{\alpha}}_{X'}^{\text{SIMEX}}$ are required, the steps (iii)–(vi) will be repeated for the parameters α_0 and $\boldsymbol{\alpha}_{X'}$.

Steps (v) and (vi) are illustrated in Figure 1 for SIMEX-Q using the example introduced in Section 6. Furthermore, it is to be noted that in step (ii) if $r_{ij}^{\lambda} + \lambda z_{ij}^{(l)}(\zeta) < 0$ for realisations of $\boldsymbol{R}_{ij}^{\lambda}$ and $Z_{ij}^{(l)}(\zeta)$, the realisation of $\boldsymbol{R}_{ij}^{(l)}(\zeta)$ will not exist for all λ. Since $\boldsymbol{R}_{ij}^{\lambda}$ is assumed to be greater than zero, this can only occur for very small values of $Z_{ij}^{(l)}(\zeta)$, which may in particular occur when the variance of $Z_{ij}^{(l)}(\zeta)$ is large. In this case, a positive constant c must be added to $\boldsymbol{R}_{ij}$ and the algorithm must be restarted with $\boldsymbol{R}_{ij} + c$ instead of $\boldsymbol{R}_{ij}$. Since this procedure is always feasible, we assume that all realisations of $\boldsymbol{R}_{ij}^{(l)}(\zeta)$ exist without loss of generality.

3.1 | Another measurement error correction: The NCI method

Here, we briefly describe the NCI method (Tooze et al., 2006; Kipnis et al., 2009) used in this paper for comparison with the SIMEX approach. The method follows a regression calibration approach under the measurement error model presented in Section 2. As in the proposed SIMEX algorithm, the non-linear measurement error model (2) is fitted first. Then, based on the estimated parameters $\hat{\lambda}, \hat{\sigma}_u^2, \hat{\sigma}_\varepsilon^2, \hat{\beta}_0$, and $\hat{\boldsymbol{\beta}}$, the usual intake T_i is estimated using the formula

$$T_i = \mathrm{E}(R_{ij}|\mathbf{X}_i, u_i) = \mathrm{E}\big(g_\lambda^{-1}(\beta_0 + \boldsymbol{\beta}\mathbf{X}_i + u_i + \varepsilon_{ij})|\mathbf{X}_i, u_i\big), \tag{7}$$

which can be approximated by a Taylor series expansion:

$$\mathrm{E}\Big(g_\lambda^{-1}(\beta_0 + \boldsymbol{\beta}\mathbf{X}_i + u_i) + \frac{1}{2}\sigma_\varepsilon^2 (g_\lambda^{-1})''(\beta_0 + \boldsymbol{\beta}\mathbf{X}_i + u_i)|\mathbf{X}_i, R_{i1}, \dots, R_{iJ_i}\Big), \tag{8}$$

where g_λ^{-1} denotes the inverse Box-Cox transformation as introduced above. This approach can be extended taking into account the consumption probabilities $P(R_{ij} > 0)$, if episodically consumed food data with excess zeros are investigated. Adaptive Gaussian Quadrature is used to estimate (8) (for details see Kipnis et al. 2009). Once the usual intake is estimated, it is plugged into in the health model (3) in place of the true usual intake T_i to derive the corrected estimate for α_T.

4 | PROPERTIES OF THE GENERATED DATA $R_{ij}^{(l)}(\zeta)$

When investigating the properties of the SIMEX algorithm, only the ME terms of all random variables are of interest. Therefore, T_i and $\mathrm{E}(g_\lambda(R_{ij}))$ are assumed to be given in what follows. The focus is on the MSE of $R_{ij}^{(l)}(\zeta)$ for each individual for $\zeta \to -1$. The question is if in this case the key property

$$\mathrm{MSE}\Big(R_{ij}^{(l)}(\zeta)\Big) \to 0 \tag{9}$$

holds. If yes, the SIMEX algorithm with the extrapolation to $\zeta = -1$ will be justified. In this hypothetical case, the generated data would be error-free for $\zeta = -1$ and the parameter estimate for these data corresponds to the parameter estimate of the unknown true data. It is important to note that according to the SIMEX algorithm, the $R_{ij}^{(l)}(\zeta)$ itself is not extrapolated but the corresponding parameter estimate of the health model is.

In Appendix A.1, using the ME model (2), it is shown that the key property (9) holds under the assumption that the corrective expected value $\mu_i(\zeta)$ guarantees $R_{ij}^{(l)}(\zeta)$ to be unbiased if $\zeta \to -1$. This assumption will be justified if the corrective expected value, is calculated using the equation

$$\mathrm{E}(R_{ij}) = (\lambda(\mathrm{E}(g_\lambda(R_{ij})) + \mu_i(\zeta)) + 1)^{\frac{1}{\lambda}} + \frac{1-\lambda}{2}(\lambda(\mathrm{E}(g_\lambda(R_{ij})) + \mu_i(\zeta)) + 1)^{\frac{1}{\lambda}-2}(1+\zeta)\sigma_\varepsilon^2 \tag{10}$$

for $\lambda > 0$ and $\mu_i(\zeta) = -\zeta\sigma_\varepsilon^2/2$ for $\lambda = 0$. This is shown in Appendix A.2. Equation (10) is well-defined if $\lambda(E(g_\lambda(R)) + \mu(\zeta)) + 1 > 0$ or equivalently $\mu(\zeta) > -((1/\lambda) + E(g_\lambda(R)))$. Furthermore, Equation (10) ensures that $\mathrm{bias}_{T_i}(R_{ij}^{(l)}(\zeta)) \approx 0$, that is, that $R_{ij}^{(l)}(\zeta)$ is a remeasurement of T_i at least approximatively. For the special cases $\lambda = 0, 1/2$, and 1, it even leads to the exact value, for example, for $\lambda = 1$ the corrective expected value simplifies to $\mu_i(\zeta) = 0$. This is also shown in Appendix A.2.

To calculate $\mu_i(\zeta)$ given the realisations r_{ij} of R_{ij}, $i = 1, \dots, I$, $j = 1, \dots, J_i$, and parameters λ, ζ, and σ_ε^2, the unknown $\mathrm{E}(R_{ij})$ and $\mathrm{E}(g_\lambda(R_{ij}))$ are replaced by the individual naïve estimates $\bar{r}_{i\bullet}$ and $\bar{r}_{i\bullet}^{(g)} = (1/J_i)\sum_1^{J_i} g_\lambda(r_{ij})$ using the empirical equation

$$\bar{r}_{i\bullet} = (\lambda(\bar{r}_{i\bullet}^{(g)} + \mu_i(\zeta)) + 1)^{\frac{1}{\lambda}} + \frac{1-\lambda}{2}(\lambda(\bar{r}_{i\bullet}^{(g)} + \mu_i(\zeta)) + 1)^{\frac{1}{\lambda}-2}\zeta\sigma_\varepsilon^2 \tag{11}$$

which can be solved numerically using an optimisation algorithm such as implemented in the R function optimise (R Core Team, 2017) (taking into account the exact corrective value for $\lambda = 0$ and $\lambda = 1$ and the well-definedness of (11), the corrective expected value should be searched in the interval $(\max\{-\zeta\sigma_\varepsilon^2/2; -((1/\lambda) + \bar{r}_{i\bullet}^{(g)})\}, 0)$ for $\lambda \in (0, 1)$). If Equation (11) cannot be satisfied, $\mu_i(\zeta)$ can still be used to ensure that the difference between $\bar{r}_{i\bullet}$ and the right-hand side is minimised. In Figure 2, the corrective expected value $\mu_i(\zeta)$ is plotted for different values of ζ and λ.

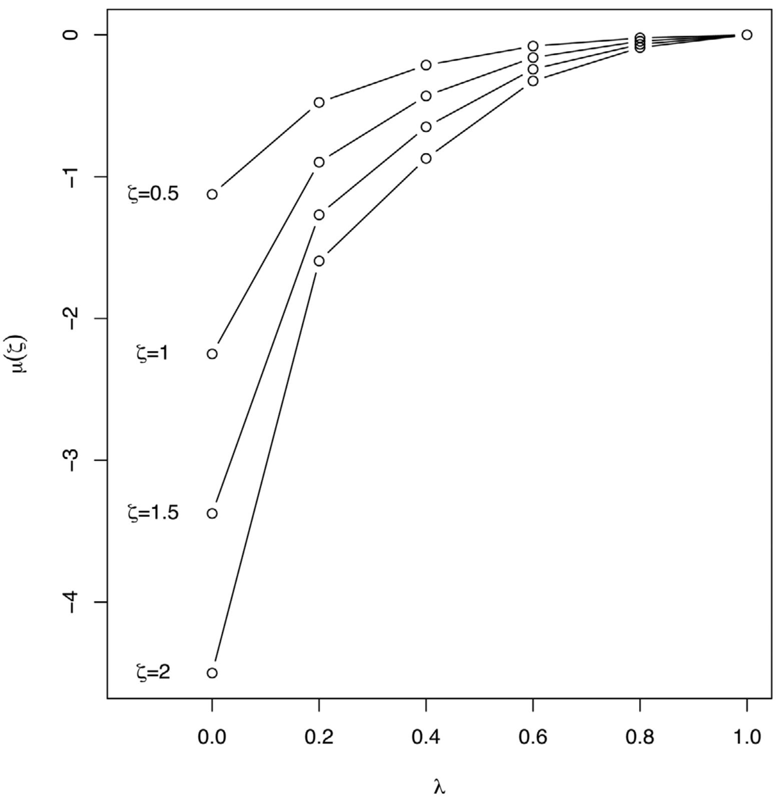

FIGURE 2 Plot of the corrective expected value $\mu(\zeta)$ depending on different values of λ and ζ: for $\zeta = 1/2, 1, 3/2, 2$; $\lambda = 1/5, 2/5, 3/5, 4/5, 1$; $\sigma_\varepsilon^2 = 4.5$; and $r = \bar{r} = 18$ calculated from the empirical Equation (11). For $\lambda = 0$ the equation $\mu(\zeta) = -\zeta\sigma_\varepsilon^2/2$ is used

5 | EXTENSION OF THE ERROR AND HEALTH MODEL

The health model is not restricted to multiple linear regression models. Depending on the outcome variable or the study design, the health model can be applied to logistic regression models or mixed models.

The error model can be extended in two different ways. The first extension will be useful for episodically consumed dietary components, such as alcohol or fish, which show a high proportion of zeros in daily consumption. According to the error model described in Kipnis et al. (2009), the true individual usual intake T_i is given by the product of the consumption probability $\mathrm{P}(T_{ij} > 0)$ and the expected intake on a consumption day $\mathrm{E}(T_{ij}|T_{ij} > 0)$:

$$T_i = \mathrm{P}(T_{ij} > 0) \times \mathrm{E}(T_{ij}|T_{ij} > 0),$$

where T_{ij} denotes the true individual daily intake. The case $R_{ij} = 0$ is allowed and occurs if and only if T_{ij} is also zero. This implies that R_{ij} is error free for T_{ij} if $R_{ij} = 0$ as $T_{ij} = R_{ij}$ when $R_{ij} = 0$ and further $\mathrm{P}(T_{ij} > 0) = \mathrm{P}(R_{ij} > 0)$. To take this into account, the first and second step of the SIMEX algorithm have to be adapted as follows.

(i) Since the error model (2) is only true for $R_{ij} > 0$, the required parameters are then estimated using only data with $R_{ij} > 0$ in the first step. This model is called the amount model.

(ii) According to the extended error model, $R_{ij}^{(l)}(\zeta)$ is defined in the second step as follows: $R_{ij}^{(l)}(\zeta) = 0$ if $R_{ij} = 0$, and $R_{ij}^{(l)}(\zeta)$ as in Equations (5) and (6) if $R_{ij} > 0$.

From steps (i) and (ii), it follows that $R_{ij}^{(l)}(\zeta) = T_{ij}$ if $R_{ij} = 0$ and also $\mathrm{P}(T_{ij} > 0) = \mathrm{P}(R_{ij}^{(l)}(\zeta) > 0)$. Given T_{ij}, the key property (9) holds for each individual i on each day j; this is obvious for $T_{ij} = 0$ since $T_{ij} = R_{ij}^{(l)}(\zeta) = 0$, and is shown in Appendix A.1 for $T_{ij} > 0$. Subsequently, in step three the full data set with $R_{ij} > 0$ and $R_{ij} = 0$ are again used to calculate the individual means $\bar{R}_{i\bullet}^{(l)}(\zeta)$. The remaining steps of the algorithm are then carried out without further modifications.

It is worth noting that in a recently proposed error model for 24HDR data, the consumption probability is modelled as in the NCI method and only the amount model is modelled differently (Agogo, 2017). For an extensive discussion of alternatives for modelling excess zeros in 24HDR data, see Kipnis et al. (2009).

If the constant term c is used to avoid the transformation error mentioned in Section 3, there are no more zeros in the data. Nevertheless, $\boldsymbol{R}_{ij} = c$ is then treated as $\boldsymbol{R}_{ij} = 0$ and the data set must be split into two subsets accordingly. When $\bar{\boldsymbol{R}}^{(l)}_{i\bullet}(\zeta)$ is calculated, c can again be subtracted though, this is not strictly necessary since the estimation of α_T is not affected.

The second extension can be used if the error variance σ^2_ε is assumed to differ between groups. For example, this could be the case if different age groups are included in the study population. Then each of the $\boldsymbol{G}$ groups is assumed to have its own error variance: $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2_{\varepsilon g})$, $g = 1, \dots, \boldsymbol{G}$. This can be easily incorporated in the first and second step of the SIMEX algorithm.

6 | APPLICATION: ASSOCIATION OF BLOOD PRESSURE WITH SALT AND ALCOHOL INTAKE

For illustrative purposes, the adapted SIMEX algorithm was applied to real data from the I.Family study, which is a multi-centre study described in detail in Ahrens et al. (2017). It is a follow-up of the IDEFICS study aiming to investigate the causes of diet- and lifestyle-related diseases in children from eight European countries (Italy, Estonia, Cyprus, Belgium, Sweden, Germany, Hungary, and Spain) (Ahrens et al., 2011). As part of the I.Family study, a web-based 24HDR was used to assess the diet of children and their parents (Hebestreit, Wolters, Jilani, Eiben, & Pala, 2018). In addition, systolic blood pressure (SBP) and the body mass index (BMI) were assessed. All institutional and governmental regulations concerning the ethical use of human volunteers were followed. Each survey centre obtained ethical approvals from the local responsible authorities in accordance with the ethical standards of the 1964 Declaration of Helsinki and its later amendments. In this study, data from 878 male adults were used to investigate the association between salt intake and blood pressure and between alcohol intake and blood pressure. The analyses were based on 1856 recalls, in which the intake of at least 500 kcal of energy was reported. The number of recalls per participant varied (38% recalled one day, 16% two days, 42% three days, and 4% four or more days). Even though no one reported zero salt consumption, the intake of less than 1 g of salt was assumed to be implausibly low. Therefore these recalls were excluded from the salt intake analysis which slightly reduces the number of included individuals and recalls to 873 and 1834.

The adapted SIMEX algorithm was applied to the salt and alcohol intake data. The covariates age, BMI, and country were used in both the ME and the health models. Since alcohol is not consumed daily (the percentage of zeros was 56%), the extended approach of Section 5 was applied to the alcohol intake data. For each value of $\zeta \in \mathcal{Z}$, the number of generated data sets was set to $\boldsymbol{L} = 200$ and $\mathcal{Z} = \{1/4, 2/4, 3/4, 1, 5/4, 6/4, 7/4, 2\}$. In the extrapolation step, the quadratic and quartic functions were used to calculate different SIMEX effect estimates. To illustrate the influence of the corrective expected value, the SIMEX data for salt intake were also generated with $\mu_i(\zeta) = 0$ for $\zeta = 2$. Furthermore, the naïve and the NCI effect estimates, using the estimated usual intakes, assuming the same health and error models, were calculated for comparison with the SIMEX effect estimates.

The variability of the estimates of α_T was assessed via the Bootstrap (Efron & Tibshirani, 1993). We drew $\boldsymbol{B} = 500$ bootstrap samples with replacement each containing the same number of individuals as the original salt and alcohol intake data set. Then, we applied the same methods as for the original data sets to the bootstrap samples and estimated the sample standard deviations of the 500 resulting effect estimates.

The estimated parameters of the error models can be found in Appendix A.3. Figure 3 shows that the inclusion of the corrective expected value resulted in a reduced bias of the generated data with regard to the true observations. The bias of the generated data using the corrective expected value was much smaller than that of the generated data without the corrective expected value.

All the different effect estimates suggest positive associations of SBP with salt and alcohol intake (Table 1). The lowest effect estimates resulted from the naïve method. Ignoring the ME led to an estimated expected increase of 0.24 mmHg in SBP per 1 gram salt intake, whereas the corrected effect estimates were 1.7–2.7 times higher than the naïve estimate. Similar results were observed for alcohol intake. Ignoring the ME led to an estimated increase of 0.339 mmHG in SBP per 10 grams alcohol intake, whereas the corrected effect estimates were 1.4–2.8 times higher than the naïve estimate. In the salt and alcohol intake analyses, the estimated standard errors were lowest for the naïve estimates and highest for the NCI estimates. The SIMEX-Q4 standard errors were comparable with those of the NCI method and 1.4–1.5 times higher than those of SIMEX-Q. That means SIMEX-Q4 and the NC -method led to the same amount of uncertainty. Although statistically significant associations could be proven in other applications using the NCI method, this was not the case for the correction methods in the applications investigated in this paper if 95% confidence intervals based on the estimated standard errors were used. Nevertheless, the correction methods led to more relevant associations from a public health perspective compared to the naïve approach and the uncertainty can be attenuated by increasing the sample size in future studies.

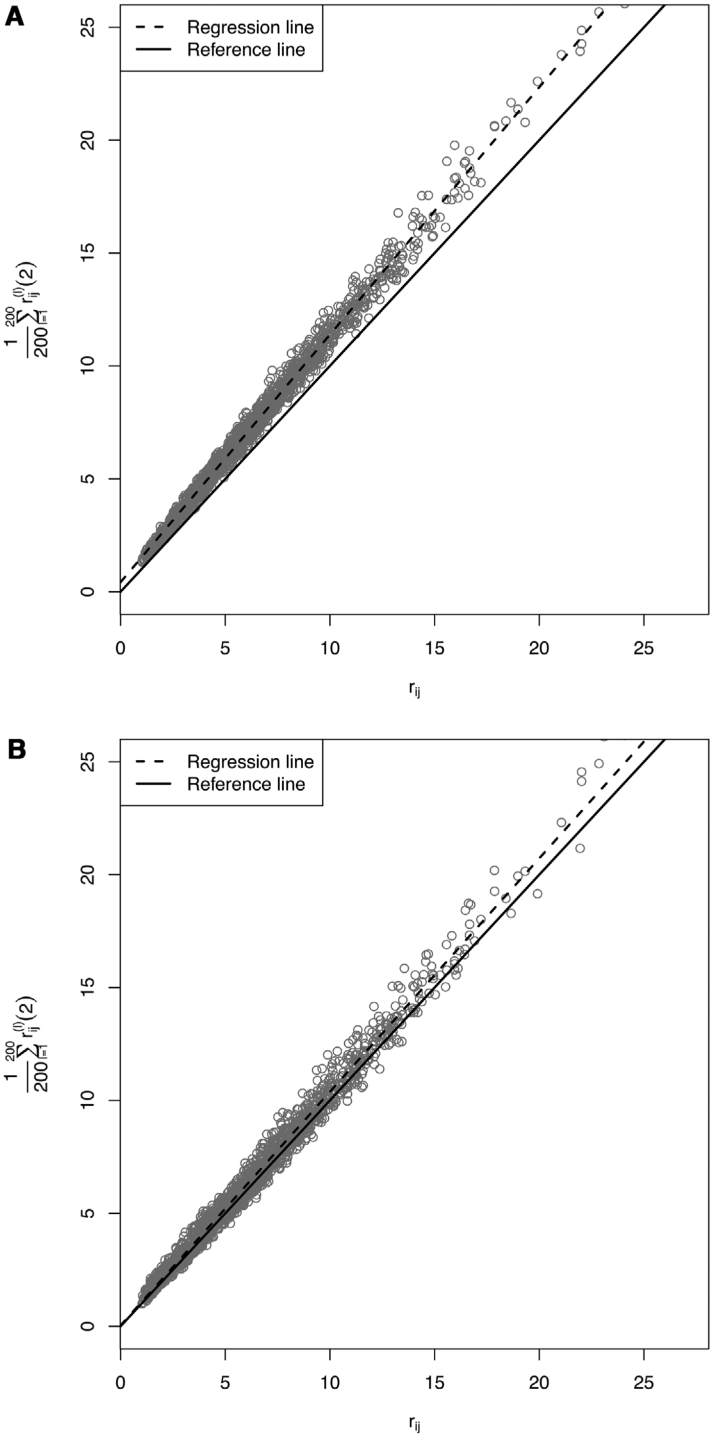

FIGURE 3 Comparison of r_{ij} and average $r_{ij}^{(l)}(\zeta)$ for $\zeta = 2$ according to the adapted SIMEX algorithm based on $L = 200$ generated SIMEX data sets (A) without bias correction using $\mu_i(\zeta) = 0$ and (B) with bias correction using $\mu_i(\zeta)$ according to Equation (11). The regression lines are $y = 0.426 + 1.096x$ and $y = 0.036 + 1.034x$

TABLE 1 Effect estimates calculated with different methods (naïve, NCI method and SIMEX with quadratic and quartic extrapolation function) and corresponding bootstrap standard errors (SE) for the association between salt intake (in grams) and the systolic blood pressure (in mmHg) and between alcohol intake (in 10 grams) and systolic blood pressure (in mmHG) in male adults adjusted for age, body mass index, and country

Exposure	**Salt intake**		**Alcohol intake**	
Method	**Estimated effect $\hat{\alpha}_T$**	**SE($\hat{\alpha}_T$)**	**Estimated effect $\hat{\alpha}_T$**	**SE($\hat{\alpha}_T$)**
Naïve	0.240	0.145	0.339	0.252
NCI	0.428	0.373	0.955	0.603
SIMEX-Q	0.404	0.237	0.474	0.378
SIMEX-Q4	0.658	0.372	0.769	0.559

TABLE 2 Comparison of effect estimates resulting from different methods (based on the true usual intake, naïve method, NCI method, and SIMEX with quadratic and quartic extrapolation functions) for two different simulation scenarios (daily and episodically consumed dietary component) with 1,000 individuals regarding empirical (emp.) mean, bias, standard error, and mean squared error based on 500 simulation data sets

Scenario	Method	Emp. mean	Emp. bias	Emp. SE	Emp. MSE
Salt intake ($\alpha_T = 0.5$)	True usual intake	0.508	0.008	0.221	0.049
	NCI	0.514	**0.014**	0.337	0.114
	SIMEX-Q4	0.422	−0.078	0.336	0.119
	SIMEX-Q	0.315	−0.185	**0.213**	**0.079**
	Naïve	0.193	−0.307	0.132	0.111
Alcohol intake ($\alpha_T = 0.9$)	True usual intake	0.887	−0.013	0.334	0.111
	NCI	0.937	**0.037**	0.532	0.284
	SIMEX-Q4	0.659	−0.241	0.472	0.280
	SIMEX-Q	0.531	−0.369	**0.315**	**0.236**
	Naïve	0.352	−0.548	0.206	0.343

7 | SIMULATION STUDY

The set-up of the simulation study was based on the studies conducted by Kipnis et al. (2009) and Agogo (2017) while the data came from the I.Family study and the models from the applications above. In total 500 data sets were simulated for each scenario and each sample size. The sample sizes were $I = 1{,}000; 500; 300; 200;$ and 100 individuals. As in the above applications, the number of recalls was assumed to vary per individual to roughly represent the distribution found in the intake data (35% with one recall, 20% with two recalls, 40% with three recalls, and 5% with four recalls), that is, the data sets consist of $2,150; 1,075; 645; 430;$ and 215 observations. For each individual, the combined covariate information for age, BMI, and country were sampled with replacement from the original data set of the application studies. For each data set and each individual, 1,000 recalls were simulated on the g_λ-transformed scale based on the error models described in Appendix A.3. These values were back-transformed and the individual average was calculated, which was used as true usual intake T_i. The same procedure (without averaging) was conducted to simulate one, two, three, or four recalls for each individual. The health outcome SBP was generated from the health models of the application studies based on the simulated true intake (for details see Appendix A.3). The true coefficients of the salt and alcohol intake α_T were chosen to be 0.5 and 0.9 (cf. Table 1) which correspond to an expected increase of 0.5 mmHg in SBP per 1 gram salt intake and of 0.9 mmHG in SBP per 10 grams alcohol intake, respectively.

The same correction methods as for the real data were applied to these simulated data sets. Furthermore, the estimates based on the true usual intake were calculated. As mentioned in Section 3, a problem occurs if $r_{ij}^{\lambda} + \lambda z_{ij}^{(l)}(\zeta) < 0$. This problem could not be completely avoided in the simulation study, since a total of $2 \times 500 \times (2,150 + 1,075 + 645 + 430 + 215) \times 8 \times 200 = 7.2 \times 10^9$ observations were generated in the SIMEX algorithm. The proportion of negative values of $r_{ij}^{\lambda} + \lambda z_{ij}^{(l)}(\zeta)$ in the generated observations ranged from 1.0×10^{-5}% to 5.1×10^{-3}% in the salt intake scenario and from 1.3% to 1.4% in the alcohol intake scenario. It was solved by using the minimum of the corresponding generated data set instead. In the alcohol intake scenario with 200 and 100 individuals, the NCI method failed for 4 and 8 data sets, respectively, due to convergence errors when fitting the non-linear mixed effects error model or failure of the Adaptive Gaussian Quadrature optimisation. These data sets were excluded from the assessment of the NCI method.

Subsequently, the performance of the different effect estimates was measured in terms of mean empirical bias, the empirical standard error (SE), and the empirical MSE. The results are summarised in Table 2 for $I = 1{,}000$ and in Table 3 for $I = 1{,}000; 500; 300; 200;$ and 100. For $I = 1{,}000$, the NCI method was nearly unbiased (0.014 and 0.037), while SIMEX-Q4 and SIMEX-Q underestimated the effects (−0.078 and −0.186 for salt intake; −0.241 and −0.369 for alcohol intake). The empirical bias of the naïve approach was most serious (−0.307 and −0.548). With respect to the empirical MSE, SIMEX-Q outperformed the other correction methods while the NCI method and SIMEX-Q4 performed equally well.

For $I = 500$, the NCI method was still superior regarding empirical bias, but if the sample size was further reduced this superiority vanished, while that of the SIMEX-Q regarding empirical MSE remained. The performance of SIMEX-Q and -Q4 regarding empirical bias seems to be independent of the sample size. SIMEX-Q4 was even less or as biased as the NCI method for $I = 300, 200,$ and 100 in the salt intake scenario and for $I = 100$ in the alcohol intake scenario. The fact that the NCI

TABLE 3 Comparison of effect estimates resulting from the NCI method and SIMEX with quadratic and quartic extrapolation functions for two different simulation scenarios (daily and episodically consumed dietary component) for different sample sizes (I = 1,000; 500; 300; 200; 100) regarding empirical (emp.) mean, bias, standard error and mean squared error based on 500 simulation data sets

Scenario	Sample size	Method	Emp. mean	Emp. bias	Emp. SE	Emp. MSE
Salt intake ($\alpha_T = 0.5$)	1,000	NCI	0.514	**0.014**	0.337	0.114
		SIMEX-Q4	0.422	−0.078	0.336	0.119
		SIMEX-Q	0.315	−0.185	**0.213**	**0.079**
	500	NCI	0.541	**0.041**	0.505	0.256
		SIMEX-Q4	0.418	−0.082	0.477	0.234
		SIMEX-Q	0.316	−0.184	**0.313**	**0.132**
	300	NCI	0.616	0.116	0.689	0.488
		SIMEX-Q4	0.453	**−0.047**	0.634	0.404
		SIMEX-Q	0.345	−0.155	**0.411**	**0.193**
	200	NCI	0.599	**0.099**	1.026	1.060
		SIMEX-Q4	0.372	−0.128	0.798	0.652
		SIMEX-Q	0.298	−0.202	**0.512**	**0.302**
	100	NCI	1.201	0.701	6.170	38.489
		SIMEX-Q4	0.428	**−0.072**	1.162	1.352
		SIMEX-Q	0.293	−0.207	**0.737**	**0.584**
Alcohol intake ($\alpha_T = 0.9$)	1,000	NCI	0.937	**0.037**	0.532	0.284
		SIMEX-Q4	0.659	−0.241	0.472	0.280
		SIMEX-Q	0.531	−0.369	**0.315**	**0.236**
	500	NCI	0.936	**0.036**	0.669	0.448
		SIMEX-Q4	0.672	−0.228	0.643	0.464
		SIMEX-Q	0.530	−0.370	**0.422**	**0.314**
	300	NCI	0.982	**0.082**	0.988	0.980
		SIMEX-Q4	0.655	−0.245	0.881	0.835
		SIMEX-Q	0.531	−0.369	**0.553**	**0.441**
	200	NCI	0.985	**0.085**	1.363	1.862
		SIMEX-Q4	0.550	−0.350	1.119	1.373
		SIMEX-Q	0.501	−0.399	**0.721**	**0.677**
	100	NCI	1.270	0.370	2.425	6.005
		SIMEX-Q4	0.738	**−0.162**	1.579	2.514
		SIMEX-Q	0.642	−0.258	**1.045**	**1.156**

method led to an overestimation of the effect in the salt intake scenario for $I = 100$ can be partly explained by seven outliers. In these cases, σ_u^2 of the ME model (2) was severely underestimated. All methods demonstrated decreasing empirical MSE with increasing sample size.

8 | DISCUSSION

In this paper, we proposed an alternative, easy-to-use method for ME correction of dietary exposure derived from a 24HDR. The method is based on the SIMEX approach and the assumptions of the error model described in Kipnis et al. (2009). We gave the justification for this method by proving the key property of the adapted SIMEX algorithm. It was shown that the MSE of the generated data converges to zero if the total ME variance converges to zero (i.e., ζ converges to -1). Furthermore, we introduced a so-called corrective expected value which ensures the generated data to be approximately unbiased on the original scale.

The underlying ME model assumes that the 24HDR is unbiased on the original scale for the true usual intake, which is not always true. Nevertheless, this is a common working assumption in this field (Dodd et al., 2006). Formally, the unbiasedness can be justified by the definition of the true usual intake as the average of daily assessed 24HDRs over a long period, since this is the best possible measure for dietary components in practice if a gold standard is not available or known (Carroll et al., 2006). Dodd et al. (2006) discussed extensively whether the unbiasedness should be assumed on the transformed or on the original scale. They are in favour of the latter. Among others, arguments for this choice are (i) that the estimated group mean usual intake and overall average of the 24HDRs coincide and (ii) that the assumption is independent of the estimated Box-Cox parameter λ, that is, it does not change by analysis group or over time. Nevertheless, if a gold standard for one specific component is available, it can be used to check the robustness of the assumption and of the proposed SIMEX algorithm and to derive new error models which may better reflect reality.

One advantage of the adapted SIMEX algorithm is the easy implementation, which is mainly based on generating the data $R_{ij}^{(l)}(\zeta), l = 1, \dots, L, \zeta \in \mathcal{Z}$, and fitting the health model to these data repeatedly. We applied the proposed method successfully to real data of daily and episodically consumed food. Furthermore, in a simulation study, we compared the proposed method with the NCI method for both scenarios and for varying sample sizes. In all scenarios, SIMEX-Q had lower empirical MSE. Although reducing bias is usually considered more important, methods aiming at a partial correction to reduce the MSE are also justified (Carroll et al., 2006). We therefore recommend applying SIMEX-Q in situations where a small MSE is considered as more important than a small bias. Otherwise, if the bias is the decisive criterion, the NCI method should be used if the sample size is sufficiently large ($I \geq 500$ or $I \geq 200$ in the investigated scenarios). If this is not the case, the SIMEX-Q4 seems to be a more attractive choice, since for smaller sample size, the method appears less biased than the NCI method. One reason for the relatively good performance of SIMEX-Q4 in these situations is the number of parameters on which SIMEX depends. In case of daily consumed food, the same ME model must be estimated for the NCI method and the adapted SIMEX algorithm, but subsequently SIMEX only uses two parameters (λ and σ_ε^2) whereas the NCI methods needs all model parameters, that is, 13 in the salt intake scenario, to estimate the usual intake. This makes the SIMEX approach more stable if one of the parameters is heavily under- or overestimated, which happened 7 out of 500 times in the salt intake scenario with $I = 100$ individuals. This unfavourable property of the NCI method is particularly relevant if the sample size is small.

The number of required parameters is also crucial when using SIMEX for sensitivity analyses. This technique, which was proposed by He, Yi, and Xiong (2007), has been recommended for situations where the ME model cannot be estimated but external information about the measurement error is available or can reasonably be assumed. This could, for example, be the case if repeated measurements are missing and values for λ and σ_ε^2 can be found in the literature. Then these parameter values can be used in the adapted SIMEX algorithm to obtain corrected effects estimates.

The proposed SIMEX modification has two drawbacks. The first is well-known from the classical SIMEX approach and was already noted by Cook and Stefanski (1994) when it was first introduced. The extrapolation is the Achilles heel of SIMEX, making it an approximative procedure (Carroll et al., 2006). That is why additional extrapolation functions were considered in the simulation study: the cubic (C) and quintic polynomial (Q5), the rational linear (RL) and spline function (Table SM1). The performance of SIMEX-C regarding empirical bias and MSE was between that of the quadratic and the quartic function, that is, inferior to SIMEX-Q4 regarding empirical bias and inferior to SIMEX-Q regarding empirical MSE. The empirical bias of SIMEX-Q5 was of similar size as the empirical bias of SIMEX-Q4, but the empirical MSE of SIMEX-Q5 was always several times higher than that of SIMEX-Q. Besides the practical problems of SIMEX-RL (see Section 3), we also observed sometimes a erratic behaviour of the rational linear extrapolation, which was already mentioned in other studies (Küchenhoff & Carroll, 1997; Carroll et al., 2006) and which resulted in an unacceptably high empirical bias and MSE in the simulation study. Instead, the spline extrapolation resulted in conservative estimates comparable to those of SIMEX-Q. From a theoretical point of view, it might also be interesting to find the exact form of the extrapolation function for the adapted SIMEX algorithm, that is, the form of the effect estimator of the health model taking into account the complex error model for 24HDRs. This will be addressed in future research. However, the practical benefit might be small considering how badly the rational extrapolation works in cases where it is actually the true form (Carroll et al., 2006).

The second drawback is due to the Box-Cox transformation in the error model in which a strictly positive variable is modelled by a normal distribution. This could mean that $R_{ij}^{(l)}(\zeta)$ cannot be calculated for some $i = 1, \dots, I$ and $j = 1, \dots, J_i$. In practice, this issue can be solved by either adding a positive constant c to R_{ij} and applying the SIMEX algorithm to $R_{ij} + c$ or if only a very few observations are affected by just setting these observations to the minimum of the observations in the specific generated data set. A theoretical solution for this could be a modified error model, for example, the truncated normal distribution. However, this makes the model more complicated and the practical benefit is low. For example, in our application on salt intake the difference between the truncated and the untruncated distribution is negligible even in the simulation step for $\zeta = 2$ since $\mathrm{P}(W_{\text{salt}} < -1/\lambda) = 10^{-11}$ if W_{salt} is conservatively estimated by $W_{\text{salt}} \sim \mathcal{N}(17, \hat{\sigma}_u^2 + 2\hat{\sigma}_\varepsilon^2)$ (as mentioned in Section 3, this negligible difference can always be achieved by adding a positive constant.). Another approach was proposed by Agogo (2017) using the generalized gamma distribution in the ME model which also increases the model complexity.

In conclusion, the proposed method is theoretically justified and led in practice to a reasonable correction. In a simulation study, the proposed method led either to estimates with smaller empirical MSE, or to estimates with smaller empirical bias for small samples than the NCI method. Furthermore, the adapted SIMEX approach can be useful in sensitivity analyses as well as graphically illustrating the measurement error correction.

ACKNOWLEDGEMENTS

The study was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 391977161. We gratefully acknowledge the financial support of the European Commission within the Seventh RTD Framework Program Contract No. 266044. We thank the I.Family children and their parents for participating in this extensive examination. We are grateful for the support from school boards, headmasters, and communities.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

ORCID

Timm Intemann https://orcid.org/0000-0001-7836-3643

REFERENCES

Agogo, G. O. (2017). A zero-augmented generalized gamma regression calibration to adjust for covariate measurement error: A case of an episodically consumed dietary intake. *Biometrical Journal*, *59*, 94–109.

Ahrens, W., Bammann, K., Siani, A., Buchecker, K., De Henauw, S., Iacoviello, L., … Pigeot, I. on behalf of the IDEFICS consortium (2011). The IDEFICS cohort: design, characteristics and participation in the baseline survey. *International Journal of Obesity*, *35*, S3–S15.

Ahrens, W., Siani, A., Adan, R., De Henauw, S., Eiben, G., Gwozdz, W., … Pigeot, I. on behalf of the I.Family consortium (2017). Cohort profile: The transition from childhood to adolescence in European children—How I.Family extends the IDEFICS cohort. *International Journal of Epidemiology: Cohort Profiles*, *46*, 1394–1395j.

Bates D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.

Boeing, H., & Margetts, B. M. (2014). Nutritional epidemiology. In W. Ahrens & I. Pigeot (Eds.), *Handbook of epidemiology* (pp. 1659–1703). New York: Springer.

Börnhorst, C., Huybrechts, I., Hebestreit, A., Krogh, V., De Decker, A., Barba, G., … Pigeot, I. on behalf of the IDEFICS consortium (2014). Usual energy and macronutrient intakes in 2–9-year-old European children. *International Journal of Obesity*, *38*, S115–S123.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: A modern perspective*. Boca Raton, FL: Chapman and Hall/CRC.

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

Dodd, K. W., Guenther, P. M., Freedman, L. S., Subar, A. F., Kipnis, V., Midthune, D., … Krebs-Smith, S. M. (2006) Statistical methods for estimating usual intake of nutrients and foods: A review of the theory. *Journal of the American Dietetic Association*, *106*, 1640–1650.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: CRC press.

Freedman, L. S., & Kipnis, V. (2018). STRengthening Analytical Thinking for Observational Studies (STRATOS): Introducing the measurement error and misclassification topic group (TG4). *Biometric Bulletin*, *35*, 10.

Freeman, J., & Modarres, R. (2006). Inverse Box-Cox: The power-normal distribution. *Statistics & Probability Letters*, *76*, 764–772.

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression*. Thousands Oaks, CA: Sage.

He, W., Yi, G. Y., & Xiong, J. (2007). Accelerated failure time models with covariates subject to measurement error. *Statistics in Medicine*, *26*, 4817–4832.

Hebestreit, A., Intemann, T., Siani, A., De Henauw, S., Eiben, G., Kourides, Y., … Pigeot, I. on behalf of the I.Family consortium (2017). Dietary patterns of European children and their parents in association with family food environment: Results from the I.Family study. *Nutrients*, *9*, 126.

Hebestreit, A., Wolters, M., Jilani, H., Eiben, G., & Pala, V. (2018). Web-based 24-hour dietary recall: The SACANA program. In K. Bammann, L. Lissner, W. Ahrens, & I. Pigeot (Eds.), *Instruments for health surveys in children and adolescents* (pp. 77–102). Heidelberg, Germany: Springer.

Intemann, T., Pigeot, I., De Henauw, S., Eiben, G., Lissner, L., Krogh, V., … Pala, V. on behalf of the I.Family consortium (2018). Urinary sucrose and fructose to validate self-reported sugar intake in children and adolescents: Results from the I.Family study. *European Journal of Nutrition*, *58*, 1247–1258. https://doi.org/10.1007/s00394-018-1649-6

Kipnis, V., Midthune, D., Buckman, D. W., Dodd, K. W., Guenther, P. M., Krebs-Smith, S. M., … Freedman, L. S. (2009). Modeling data with excess zeros and measurement error: Application to evaluating relationships between episodically consumed foods and health outcomes. *Biometrics*, *65*, 1003–1010.

Küchenhoff, H., & Carroll, R. J. (1997). Segmented regression with errors in predictors: Semi-parametric and parametric methods. *Statistics in Medicine*, *16*, 169–188.

Lederer, W., & Küchenhoff, H. (2006). A short introduction to the SIMEX and MCSIMEX. *R News*, *6*, 26–31.

Liese, A. D., Crandell, J. L., Tooze, J. A., Kipnis, V., Bell, R., Couch, S. C., … Mayer-Davis, E. J. (2015). Sugar-sweetened beverage intake and cardiovascular risk factor profile in youth with type 1 diabetes: Application of measurement error methodology in the SEARCH Nutrition Ancillary Study. *British Journal of Nutrition*, *114*, 430–438.

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Shaw, P. A., Deffner, V., Keogh, R. H., Tooze, J. A., Dodd, K. W., Küchenhoff, H., … Freedman, L. S. (2018). Epidemiologic analyses with error-prone exposures: Review of current practice and recommendations. *Annals of Epidemiology*, *28*, 821–828.

Souverein, O. W., Dekkers, A. L., Geelen, A., Haubrock, J., de Vries, J. H., Ocké, M. C., … van't Veer, P. (2011). Comparing four methods to estimate usual intake distributions. *European Journal of Clinical Nutrition*, *65*, S92–S101.

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

Tooze, J. A., Midthune, D., Dodd, K. W., Freedman, L. S., Krebs-Smith, S. M., Subar, A. F., … Kipnis, V. (2006). A new statistical method for estimating the usual intake of episodically consumed foods with application to their distribution. *Journal of the American Dietetic Association*, *106*, 1575–1587.

SUPPORTING INFORMATION

Additional supporting information including source code to reproduce the results may be found online in the Supporting Information section at the end of the article.

How to cite this article: Intemann T, Mehlig K, De Henauw S, Siani A, Constantinou T, Moreno LA, Molnár D, Veidebaum T, Pigeot I. SIMEX for correction of dietary exposure effects with Box-Cox transformed data. *Biometrical Journal*. 2020;62:221–237. https://doi.org/10.1002/bimj.201900066

APPENDIX

A.1 Proof of the key property

For sake of clarity, we will omit the indices i and j below. We assume that T and $\mathrm{E}(g_\lambda(R))$ are given and define a random variable W as $W = \mathrm{E}(g_\lambda(R)) + \varepsilon + Z^{(l)}(\zeta)$ which follows a normal distribution with parameters $\mu_W = \mathrm{E}(g_\lambda(R)) + \mu(\zeta)$ and $\sigma_W^2 = (1+\zeta)\sigma_\varepsilon^2$. According to (1) and (5), it follows that $R^{(l)}(\zeta) = g_\lambda^{-1}(W)$.

To show that the key property holds (under the assumption that $R^{(l)}(\zeta)$ is unbiased for each individual), we first consider the case $\lambda = 0$. Then, $R^{(l)}(\zeta) = g_\lambda^{-1}(W) = \exp(W)$ which is logarithmically normally distributed. Therefore, it follows

$$\mathrm{Var}(R^{(l)}(\zeta)) = [\exp(\sigma_W^2) - 1]\exp(2\mu_W + \sigma_W^2) = [\exp((1+\zeta)\sigma_\varepsilon^2) - 1]\exp(2\mu_W + (1+\zeta)\sigma_\varepsilon^2).$$

If $\zeta \to -1$, the first factor will converge to zero and the second to a constant and thus $\lim_{\zeta\to -1}\mathrm{Var}(R^{(l)}(\zeta)) = 0$.

Now, we consider the case $\lambda > 0$. Then the density function of $g_\lambda^{-1}(W)$ is

$$f_{R^{(l)}}(x) = \frac{1}{\sqrt{2\pi\sigma_W^2}}x^{\lambda-1}\exp\left(-\frac{(g_\lambda(x)-\mu_W)^2}{2\sigma_W^2}\right).$$

We consider the second moment of $R^{(l)}(\zeta)$ using the substitution $x = g_\lambda^{-1}(\varphi)$, the inequality $(\lambda x + 1) \leq \exp(\lambda x)$ and again the variance of the logarithmic normal distribution:

$$\mathrm{E}(R^{(l)}(\zeta)^2) = \int_0^\infty x^2 f_{R^{(l)}}(x)dx$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{2\pi\sigma_W^2}} \int_0^\infty x^{\lambda+1} \exp\left(-\frac{(g_\lambda(x)-\mu_W)^2}{2\sigma_W^2}\right) dx \\
&= \frac{1}{\sqrt{2\pi\sigma_W^2}} \int_{-\frac{1}{\lambda}}^\infty (\lambda\varphi+1)^{\frac{2}{\lambda}} \exp\left(-\frac{(\varphi-\mu_W)^2}{2\sigma_W^2}\right) d\varphi \\
&\le \frac{1}{\sqrt{2\pi\sigma_W^2}} \int_{-\frac{1}{\lambda}}^\infty \exp(\varphi)^2 \exp\left(-\frac{(\varphi-\mu_W)^2}{2\sigma_W^2}\right) d\varphi \\
&\le \frac{1}{\sqrt{2\pi\sigma_W^2}} \int_{-\infty}^\infty \exp(\varphi)^2 \exp\left(-\frac{(\varphi-\mu_W)^2}{2\sigma_W^2}\right) d\varphi \\
&= \exp(2\mu_W + 2\sigma_W^2). \qquad \text{(A.1)}
\end{aligned}
$$

It follows the existence of the second moment and the existence of the expected value and variance of $\boldsymbol{R}^{(l)}(\zeta)$.

Freeman and Modarres (2006) found an alternative representation for the moments of $g_\lambda^{-1}(W)$ using the Taylor expansion for $(g_\lambda^{-1})^r$ at μ_W:

$$
\mathrm{E}(\boldsymbol{R}^{(l)}(\zeta)^r) = (\lambda\mu_W+1)^{\frac{r}{\lambda}} + \sum_{\iota=1}^{\infty} \frac{1}{\iota!}(\lambda\mu_W+1)^{\frac{r}{\lambda}-\iota}\sigma_W^\iota \mathrm{E}(\Psi^\iota)\prod_{j=0}^{\iota-1}(r-j\lambda), \qquad \text{(A.2)}
$$

where Ψ denotes the standardized random variable $\Psi = (W-\mu_W)/\sigma_W$. Thus the variance has the following form:

$$
\begin{aligned}
&\mathrm{Var}(\boldsymbol{R}^{(l)}(\zeta)) \\
&= \mathrm{E}(\boldsymbol{R}^{(l)}(\zeta)^2) - \left(\mathrm{E}(\boldsymbol{R}^{(l)}(\zeta))\right)^2 \\
&= (\lambda\mu_W+1)^{\frac{2}{\lambda}} + \sum_{\iota=1}^{\infty}\frac{1}{\iota!}(\lambda\mu_W+1)^{\frac{2}{\lambda}-\iota}(1+\zeta)^{\frac{\iota}{2}}\sigma^\iota \mathrm{E}(\Psi^\iota)\prod_{j=0}^{\iota-1}(2-j\lambda) \\
&- \left((\lambda\mu_W+1)^{\frac{1}{\lambda}} + \sum_{\iota=1}^{\infty}\frac{1}{\iota!}(\lambda\mu_W+1)^{\frac{1}{\lambda}-\iota}(1+\zeta)^{\frac{\iota}{2}}\sigma^\iota \mathrm{E}(\Psi^\iota)\prod_{j=0}^{\iota-1}(1-j\lambda)\right)^2 \\
&= (\lambda\mu_W+1)^{\frac{2}{\lambda}} + \sum_{\iota=1}^{\infty}\frac{1}{\iota!}(\lambda\mu_W+1)^{\frac{2}{\lambda}-\iota}(1+\zeta)^{\frac{\iota}{2}}\sigma^\iota \mathrm{E}(\Psi^\iota)\prod_{j=0}^{\iota-1}(2-j\lambda) \\
&- (\lambda\mu_W+1)^{\frac{2}{\lambda}} \\
&- 2(\lambda\mu_W+1)^{\frac{1}{\lambda}}\sum_{\iota=1}^{\infty}\frac{1}{\iota!}(\lambda\mu_W+1)^{\frac{1}{\lambda}-\iota}(1+\zeta)^{\frac{\iota}{2}}\sigma^\iota \mathrm{E}(\Psi^\iota)\prod_{j=0}^{\iota-1}(1-j\lambda) \\
&- \left(\sum_{\iota=1}^{\infty}\frac{1}{\iota!}(\lambda\mu_W+1)^{\frac{1}{\lambda}-\iota}(1+\zeta)^{\frac{\iota}{2}}\sigma^\iota \mathrm{E}(\Psi^\iota)\prod_{j=0}^{\iota-1}(1-j\lambda)\right)^2 \\
&= \sum_{\iota=1}^{\infty}\frac{1}{\iota!}(\lambda\mu_W+1)^{\frac{2}{\lambda}-\iota}(1+\zeta)^{\frac{\iota}{2}}\sigma^\iota \mathrm{E}(\Psi^\iota)\prod_{j=0}^{\iota-1}(2-j\lambda) \qquad \text{(A.3)} \\
&-2(\lambda\mu_W+1)^{\frac{1}{\lambda}}\sum_{\iota=1}^{\infty}\frac{1}{\iota!}(\lambda\mu_W+1)^{\frac{1}{\lambda}-\iota}(1+\zeta)^{\frac{\iota}{2}}\sigma^\iota \mathrm{E}(\Psi^\iota)\prod_{j=0}^{\iota-1}(1-j\lambda) \qquad \text{(A.4)}
\end{aligned}
$$

$$-\left(\sum_{\iota=1}^{\infty}\frac{1}{\iota!}(\lambda\mu_W+1)^{\frac{1}{\lambda}-\iota}(1+\zeta)^{\frac{\iota}{2}}\sigma^{\iota}\mathrm{E}(\Psi^{\iota})\prod_{j=0}^{\iota-1}(1-j\lambda)\right)^2. \tag{A.5}$$

As shown before in (A.1), the infinite sums (A.3), (A.4), and (A.5) converge. Thus, considering (A.3) and $\zeta \to -1$, it follows:

$$\begin{aligned}
&\lim_{\zeta\to-1}\sum_{\iota=1}^{\infty}\frac{1}{\iota!}(\lambda\mu_W+1)^{\frac{2}{\lambda}-\iota}(1+\zeta)^{\frac{\iota}{2}}\sigma^{\iota}\mathrm{E}(\Psi^{\iota})\prod_{j=0}^{\iota-1}(2-j\lambda)\\
&=\lim_{\zeta\to-1}\lim_{n\to\infty}\sum_{\iota=1}^{n}\frac{1}{\iota!}(\lambda\mu_W+1)^{\frac{2}{\lambda}-\iota}(1+\zeta)^{\frac{\iota}{2}}\sigma^{\iota}\mathrm{E}(\Psi^{\iota})\prod_{j=0}^{\iota-1}(2-j\lambda)\\
&=\lim_{n\to\infty}\lim_{\zeta\to-1}\sum_{\iota=1}^{n}\frac{1}{\iota!}(\lambda\mu_W+1)^{\frac{2}{\lambda}-\iota}(1+\zeta)^{\frac{\iota}{2}}\sigma^{\iota}\mathrm{E}(\Psi^{\iota})\prod_{j=0}^{\iota-1}(2-j\lambda)\\
&=\lim_{n\to\infty}\sum_{\iota=1}^{n}\frac{1}{\iota!}(\lambda\mu_W+1)^{\frac{2}{\lambda}-\iota}\cdot 0\cdot\sigma^{\iota}\mathrm{E}(\Psi^{\iota})\prod_{j=0}^{\iota-1}(2-j\lambda)\\
&=0.
\end{aligned}\tag{A.6}$$

The same argument applies also for the infinite sums (A.4) and (A.5). Therefore, $\lim_{\zeta\to-1}\mathrm{Var}(\boldsymbol{R}^{(l)}(\zeta))=0$ for any μ_W.

In the important canonical case $\lambda=1$, that is, $g_\lambda(v)=v-1$, it simplifies to $\boldsymbol{R}^{(l)}(\zeta)=\mathrm{E}(\boldsymbol{R})+\varepsilon+Z^{(l)}(\zeta)$ and therefore $\mathrm{Var}(\boldsymbol{R}^{(l)}(\zeta))=(1+\zeta)\sigma^2$. This converges to zero for all $\mathrm{E}(\boldsymbol{R})$.

A.2 Calculation of the corrective expected value $\mu_i(\zeta)$ and proofs of implications

In general $\boldsymbol{R}^{(l)}(\zeta)$ given T and $\mathrm{E}(g_\lambda(\boldsymbol{R}))$ for each individual will be biased if $\zeta\to-1$. Therefore, a specific so-called corrective expected value $\mu(\zeta)$ is calculated in order to ensure unbiasedness. In the case $\lambda=0$, $g_\lambda=\log$, the corrective expected value is $-\zeta\sigma_\varepsilon^2/2$ (cf. Carroll et al. 2006), since using the expected value of the log-normal distribution gives

$$\begin{aligned}
\mathrm{E}\big(\boldsymbol{R}^{(l)}(\zeta)\big)&=\mathrm{E}\big(\exp\big(\log(\boldsymbol{R})+Z^{(l)}(\zeta)\big)\big)\\
&=\mathrm{E}(\boldsymbol{R})\mathrm{E}\big(\exp\big(Z^{(l)}(\zeta)\big)\big)\\
&=\mathrm{E}(\boldsymbol{R})\exp\left(\mu(\zeta)+\frac{\zeta\sigma_\varepsilon^2}{2}\right)\\
&=\mathrm{E}(\boldsymbol{R}).
\end{aligned}$$

In order to calculate the corrective expected value to ensure $\mathrm{E}(\boldsymbol{R}^{(l)}(\zeta))=\mathrm{E}(\boldsymbol{R})$ at least approximately, we use the Taylor approximation for $\mathrm{E}(g_\lambda^{-1}(W))$ at μ_W:

$$\begin{aligned}
\mathrm{E}(g_\lambda^{-1}(W))&\approx\mathrm{E}\left(g_\lambda^{-1}(\mu_W)+(g_\lambda^{-1})'(\mu_W)(W-\mu_W)+\frac{(g_\lambda^{-1})''(\mu_W)}{2}(W-\mu_W)^2\right)\\
&=g_\lambda^{-1}(\mu_W)+\frac{(g_\lambda^{-1})''(\mu_W)}{2}(1+\zeta)\sigma_\varepsilon^2\\
&=(\lambda\mu_W+1)^{\frac{1}{\lambda}}+\frac{1-\lambda}{2}(\lambda\mu_W+1)^{\frac{1}{\lambda}-2}(1+\zeta)\sigma_\varepsilon^2\\
&=(\lambda(E(g_\lambda(\boldsymbol{R}))+\mu(\zeta))+1)^{\frac{1}{\lambda}}\\
&\quad+\frac{1-\lambda}{2}(\lambda(E(g_\lambda(\boldsymbol{R}))+\mu(\zeta))+1)^{\frac{1}{\lambda}-2}(1+\zeta)\sigma_\varepsilon^2.
\end{aligned}$$

Therefore, Equation (10) is used to calculate $\mu(\zeta)$. In the following, we show desirable properties for the three special cases, $\zeta \to -1$, $\lambda = 1$, and $\lambda = 1/2$ using this equation.

In the first case, on the one hand Equation (10) simplifies to $E(R) = (\lambda(E(g_\lambda(R)) + \mu(-1)) + 1)^{\frac{1}{\lambda}} = g_\lambda^{-1}(E(g_\lambda(R)) + \mu(-1))$. It follows that $\mu(-1) = g_\lambda(E(R)) - E(g_\lambda(R))$. On the other hand, if $\zeta \to -1$ and using Equation (A.2) and the same argumentation as in (A.6) it follows that $\mathrm{E}(\boldsymbol{R}^{(l)}(\zeta)) = (\lambda\mu_W + 1)^{\frac{1}{\lambda}}$. Using the definition of μ_W, it follows $\mathrm{E}(\boldsymbol{R}^{(l)}(\zeta)) = (\lambda(g_\lambda(\mathrm{E}(\boldsymbol{R}))) + 1)^{\frac{1}{\lambda}} = g_\lambda^{-1}(g_\lambda((\mathrm{E}(\boldsymbol{R}))) = \mathrm{E}(\boldsymbol{R})$.

In the second case, $\lambda = 1$, $g_1(v) = v - 1$, Equation (10) simplifies to $\mu(\zeta) = g_1(E(R)) - E(g_1(R)) = \mathrm{E}(R) - 1 - \mathrm{E}(R - 1) = 0$. Obviously, this is the exact solution for $\mu(\zeta)$ to ensure $\mathrm{E}(\boldsymbol{R}^{(l)}(\zeta)) = \mathrm{E}(\boldsymbol{R})$.

In the third case, $\lambda = 1/2$, on the one hand Equation (10) simplifies to $E(R) = (0.5(E(g_{0.5}(R)) + \mu(\zeta)) + 1)^2 + 0.25(1 + \zeta)\sigma_\varepsilon^2$ which is equivalent to $\mu(\zeta) = \pm 2\sqrt{E(R) - 0.25(1 + \zeta)\sigma_\varepsilon^2} - 2 - E(g_{0.5}(R))$. On the other hand, $\mathrm{E}(\boldsymbol{R}^{(l)}(\zeta)) = \mathrm{E}(g_{0.5}^{-1}(W))$ with $g_{0.5}^{-1}(v) = (0.5v + 1)^2$ simplifies to $\mathrm{E}(0.25W^2 + W + 1)$ and using the first and second moment of the normal distribution will lead to $\mu(\zeta)$ as calculated before if $\mathrm{E}(\boldsymbol{R}^{(l)}(\zeta)) = \mathrm{E}(\boldsymbol{R})$.

A.3 Estimated measurement error and health models

To estimate the non-linear mixed model (2), the R-functions lmer and the powerTransform of the R-packages lme4 (Fox & Weisberg, 2011) and car (Bates, Mächler, Bolker, & Walker 2015) were used in the SIMEX algorithm. In the salt intake analysis the estimated parameters were

$$\begin{aligned}
\hat{\lambda} &= 0.197,\\
\hat{\sigma}_u^2 &= 1.781,\\
\hat{\sigma}_\varepsilon^2 &= 4.526,\\
\hat{\beta}_0 &= 18.349 \text{ and}\\
\hat{\boldsymbol{\beta}} &= (-0.016, 0.015, -0.341, -0.521, 0.938, 0.791, -0.061, 1.055, -0.255),
\end{aligned}$$

where $\hat{\boldsymbol{\beta}}$ includes the parameter estimates of the covariates age, BMI and the countries Estonia, Cyprus, Belgium, Sweden, Germany, Hungary, and Spain (Italy serves as reference category.). To estimate the error model for the NCI methods, the SAS-macros mixtran and indivint were used (Kipnis et al., 2009) resulting in slightly different parameter estimates, for example, $\hat{\lambda} = 0.2005$. The estimated parameters of the health model using the usual salt intake calculated with the NCI method were:

$$\begin{aligned}
\hat{\alpha}_0 &= 88.28\\
\hat{\boldsymbol{\alpha}}_X &= (0.18, 0.92, 1.09, 0.36, -3.75, -0.97, 2.36, 3.35, 2.47) \text{ and}\\
\hat{\sigma}_H^2 &= 143.12,
\end{aligned}$$

where σ_H^2 is the variance of the error term of health model. The parameter vector $\boldsymbol{\alpha}_X$ has the same structure as $\boldsymbol{\beta}$.

Analogously, for the alcohol intake analysis, the parameters in the amount model, that is, for $\boldsymbol{R}_{ij} > 0$, were estimated as follows using the SAS-macro mixtran:

$$\begin{aligned}
\hat{\lambda} &= 0.314,\\
\hat{\sigma}_u^2 &= 3.634,\\
\hat{\sigma}_\varepsilon^2 &= 5.931,\\
\hat{\beta}_0 &= 5.409 \text{ and}\\
\hat{\boldsymbol{\beta}} &= (0.005, -0.045, -0.429, -1.111, 0.766, 0.309, -1.961, -2.442, -1.939).
\end{aligned}$$

Again, the estimates obtained with the R-functions for the SIMEX algorithm only slightly differed (e.g., $\hat{\lambda} = 0.313$).

For the NCI method, the probability of the (reported) alcohol intake needed to be estimated using the mixed effects logistic regression model:

$$P(R_{ij} > 0|X_i, u_i') = \text{logit}^{-1}\left(\beta_0' + \boldsymbol{\beta}'\mathbf{X}_i + u_i'\right),$$

with the parameters β_0' and $\boldsymbol{\beta}'$ and the random variable $u_i' \sim \mathcal{N}(0, \sigma^2_{u'})$ (X_i as before). Furthermore, the correlation ρ between u_i' and u_i (of model (2)) needed to be estimated to allow the probability and the amount of intake to be correlated (for details see Kipnis et al. 2009). The estimated parameters were:

$$\hat{\sigma}^2_{u'} = 2.048,$$

$$\hat{\beta}_0' = 0.747,$$

$$\hat{\boldsymbol{\beta}}' = (0.048, -0.077, -0.145, -0.479, 0.446, -1.231, -0.031, -0.720, 0.874) \text{ and}$$

$$\hat{\rho} = 0.348.$$

The estimated parameters of the corresponding health model using the usual alcohol intake calculated with the NCI method were:

$$\hat{\alpha}_0 = 89.65$$

$$\hat{\boldsymbol{\alpha}}_X = (0.16, 0.96, 1.13, 0.52, -3.85, -0.22, 2.75, 4.73, 2.63) \text{ and}$$

$$\hat{\sigma}^2_H = 143.30.$$