COMBINING BIOMARKER AND FOOD INTAKE DATA: CALIBRATION EQUATIONS FOR CITRUS INTAKE

Silvia D'Angelo[1,2], Isobel Claire Gormley[2], Breige A McNulty[1], Anne P Nugent[3], Janette Walton[4, 5], Albert Flynn[4], and Lorraine Brennan[1].

Names for PubMed indexing: D'Angelo, Gormley, McNulty, Nugent, Walton, Flynn, Brennan.

[1] Institute of Food and Health, School of Agriculture and Food Science, University College Dublin, Dublin, Ireland

[2] School of Mathematics and Statistics, Insight Centre for Data Analytics, University College Dublin, Dublin, Ireland

[3] Institute for Global Food Security, School of Biological Sciences, Queens University Belfast, Northern Ireland.

[4] School of Food and Nutritional Sciences, University College Cork, Cork, Ireland

[5] Dept. Biological Sciences, Cork Institute of Technology, Cork, Ireland

Address correspondence to L Brennan, Institute of Food and Health, Belfield, University College Dublin, Dublin 4, Ireland. E-mail: lorraine.brennan@ucd.ie ; Phone: 00 353 1 7162811

Supported by a research grant from The European Research Council ERC (647783) and from a Science Foundation Ireland grant (SFI/12/RC/2289).

Running heading: Calibration equations for citrus intake

Abbreviations: MSE (Mean Squared Error), R^2 (R squared).

Abstract

Background

Measurement error associated with self-reported dietary intake is a well-documented issue. Combining biomarkers of food intake and dietary intake data is a high priority.

Objectives

The objective was to develop calibration equations for food intake, illustrated with an application on citrus intake. Further, a simulation-based framework was developed to determine the portion of biomarker data needed for stable calibration equations estimation in large population studies.

Design

Calibration equations were developed using mean daily self-reported citrus intake (4 days semi-weighed food diaries) and biomarker-derived intake (urinary proline betaine biomarker) data from participants (n=565) as part of a cross-sectional study. Different functional specifications and biomarker transformations were tested to derive the optimal calibration equations specification. The simulation study was developed using linear regression for the calibration equations. Stability in the calibration equations estimation was investigated for varying portions of biomarker and intake data "qualities".

Results

With citrus intake, linear regression on non-transformed biomarker data resulted in the optimal calibration equations specification and produced good quality predicted intakes. The lowest MSE (14354) corresponded to a linear regression model, defined with biomarker-derived estimates of intakes on the original scale. Using this model in a sub-population without biomarker data resulted in an average mean citrus intake of 81±66 g/day. The simulation study suggested that in large population studies, biomarker data on 20-30% of the

subjects are required to guarantee stable estimation of calibration equations. The paper is accompanied by a web application (“Bio-Intake”), developed to facilitate measurement error correction in self-reported mean daily citrus intake data.

Conclusion

Calibration equations proved to be a useful instrument to correct measurement error in self-reported food intake data. The simulation study demonstrated that the use of food intake biomarkers may be feasible and beneficial in the context of large population studies.

Keywords

Biomarkers, calibration equations, measurement error, citrus, proline betaine

INTRODUCTION

Problems surrounding our ability to accurately measure dietary intake have been well documented in recent years in the literature (1-3). To overcome the measurement error associated with self-reported dietary intake the concept of dietary biomarkers has emerged. These biomarkers can deliver objective measures of intake and examples include urinary sodium and nitrogen for salt and protein intake respectively (1). Furthermore, there is evidence that use of biomarkers to develop calibrated equations to correct self-reported intake for estimates for protein and energy intake can positively impact on the ability to examine diet-disease associations: for example inclusion of biomarker calibrated data in the Women's Health Initiative cohorts allowed for disease associations to be revealed that otherwise would not have been identified (4, 5). Using urinary nitrogen and doubly labelled water as biomarkers to calibrate intakes of protein and energy in a group of postmenopausal women revealed associations between protein and energy intake with diabetes risk. It is noteworthy that these associations were not identified in the uncalibrated data.

The majority of the work to date in this field has focused on energy and protein intake using urinary nitrogen and doubly labelled water as recovery biomarkers. However, recent data has demonstrated that a series of biomarkers including carotenoids, tocopherols, folate, vitamin B12, and phospholipid fatty acids performed as well as established energy (doubly labelled water) and protein (urinary nitrogen) biomarkers in representing nutrient intake (6). This highlights the potential of widening the nutrients that could be corrected for measurement error using biomarkers and demonstrates that estimating intakes is not limited to recovery biomarkers. Furthermore, data emerging from the application of metabolomics to biomarker discovery has revealed that food intake biomarkers may be used to estimate food intake opening up the possibility of development of calibration equations for intake of foods (7-9).

The objective of the present work was to develop calibration equations for citrus intake using urinary proline betaine as the biomarker. Furthermore, we demonstrate through a simulation study the required portion of biomarker measures necessary to enable development of calibration equations in a large population group thus laying the foundations for future work. Finally, a web application ("Bio-Intake") was developed to perform correction of measurement error in self-reported mean daily citrus intake data, based on the findings of this paper (https://adiet.shinyapps.io/Bio-Intake.)

METHODS

STATISTICAL ANALYSIS

Development of calibration equations for estimation of citrus intake

All data used for the calibration equations was obtained from previously reported studies (7). The selection process for the data is illustrated through a flow chart in the supplementary material (**Supplementary Figure 1**) and the key variables are self-reported citrus intake and urinary proline betaine. Briefly, dietary data was collected, over four consecutive days, using a four day semi-weighed food diary. Mean daily citrus intake (average citrus intake based on the four days of recording) was computed for the total citrus food group. The biomarker proline betaine has been previously presented a good biomarker of citrus intake with correlations of 0.9 with actual citrus intake. Furthermore, this has been demonstrated in different study settings by multiple research teams (9-12). Our previous work indicated that urinary proline betaine levels accurately estimated total citrus intake in a free living population group (7).

The calibration equations method relies on a central assumption, that both self-reported and biomarker-derived estimates of intake of study participant *i* are linearly related to the so-

called “true” intake X_i, with the relations being perturbated by Gaussian errors. However, as true intakes are unknown, directly estimating such relationships is not possible. The calibration method overcomes this issue by deriving a conditional prediction of X_i, $\widehat{X}_i$, given the biomarker-derived estimates of intake M_i (13). Such conditional predictions are derived from the calibrated intakes W_i, i.e. the self-reported intakes adjusted for measurement error:

$$M_i = \beta_0 + \beta_1 W_i + \epsilon_i \quad ; \quad \widehat{X}_i = \widehat{\beta_0} + \widehat{\beta_1} W_i$$

The data analyzed herein refer to a study conducted on $N = 1500$ subjects, indexed by $i = 1, \dots, N$ and collectively denoted A_N, where food consumption data was recorded using 4 day semi weighed food diaries. Mean daily citrus intake W_i (g/day) was computed for each subject. Also, biomarker-derived estimates of citrus intake M_i were computed for a subsample of A_N, here denoted A_n, with $n = 565$. Self-reported data and biomarker estimated intake data for these n subjects were used to develop calibration equations. The estimated calibration equations were then employed to correct self-reported mean daily citrus intake W_i for the remaining $(N - n) = n^* = 935$ subjects, collectively denoted A_{n*}.

The supplementary material provides a more detailed description of the calibration method. Common problems that may be encountered with food intake data include zero inflation (14), due the presence of non-consumers, and non-Gaussianity of the errors ϵ_i. Potentially, such issues could result in poor correction of measurement error. In the literature, possible non-Gaussianity of the errors has been addressed using different transformations of the biomarker-derived estimates of intake, M_i, such as the log transformation (15) or Box-Cox transformations (16). In previous work (16), zero-inflated data have been commonly modeled with the assistance of covariates. However, when such subject-specific information is not available, such an approach cannot be pursued.

In the present work, transformations of the biomarker-derived intake data have been considered for the estimation of calibration equations, as well as original scale biomarker-derived estimates of intake, to address potential non-Gaussianity of subject-specific errors. Additionally, to address the issue of zero-inflated data in the absence of covariates, zero-inflated models have been explored and compared with a standard linear regression model, to select the optimal model specification for the calibration equations. The average mean squared error (MSE) between biomarker-derived estimates of intake and calibrated mean-daily self-reported intake has been used to compare different model specifications and different transformations of the biomarker-derived intake data. The supplement presents more detail on the different models explored and gives a thorough discussion of the results obtained.

Estimation of the required portion of biomarker data in large studies

Acquisition of biomarker derived estimates of intake for all subjects may not be feasible in large population studies, due to high costs and limited biological samples availability. In order to estimate the required quantity of biomarker data needed in large studies a simulation study has been conducted. A set of $\mathrm{P} = 100$ datasets of $\mathrm{N} = 100{,}000$ subjects was created. For each dataset p $(\mathrm{p} = 1, \dots, \mathrm{P})$ biomarker-derived estimates of intake and self-reported intake data have been simulated, according to the results obtained from the analysis of mean-daily citrus intake data. Specifically, true mean-daily citrus intake values (g/day) have been simulated from a Gaussian distribution, with mean and variance fixed as the empirical mean and variance of the calibrated mean-daily self-reported citrus intakes. Self-reported and biomarker-derived intakes (g/day) have been then constructed from the simulated true intakes, following a classical calibration equations framework (17). To represent different situations that could be encountered with real data, simulated self-reported and biomarker

intakes may either have moderate or high variability around the "true intake". Such variation is regulated by the parameters α_W and α_M for self-reported and biomarker-derived intakes respectively. Also, different strengths of the association between self-reported and true intakes have been accounted for, to represent different self-reported data qualities. This feature is regulated by the parameter β_W; further detail on parameters $\alpha_{W,}\alpha_M$ and β_W are given in the supplement. Biomarker derived intakes are assumed to have been measured only for n_g of the N subjects, with $n_g = gN$ and $g = (0.01, 0.02, \dots, 0.99,1)$. For each value of n_g, simulated intakes were used to estimate calibration equations, with the corresponding parameters being estimated via cross-validation (100 repetitions, excluding 20% of the observations at each round). Standard deviations of the estimated parameters, $sd(\widehat{\beta_0})$ and $sd(\widehat{\beta_1})$, were computed and compared for different values of n_g, i.e. for different sizes of the sub-sample of subjects for which biomarker data are available. A stabilization of the parameter estimates indicates that, after a given n_g, further increasing the number of subjects for which biomarker-derived intakes are computed would not yield a substantial improvement in measurement error correction. More details regarding the construction of the simulation study are given in the supplement, while the results are presented in the next section.

RESULTS

Estimating citrus intake using calibration equations

Summary statistics for self-reported and biomarker-derived mean daily citrus intakes (g/day) are reported in

Table *1*. All examined measures of citrus intake have similar interquartile ranges and similar means, $\overline{W_{A_N}} = 75.42$ g/day and $\overline{M_{A_n}} = 83.26$ g/day. Biomarker-derived estimates of

intake range in a wider interval ($\mathrm{Max}\left(\mathrm{M}_{\mathrm{A_n}}\right) = 1226.79$) compared to self-reported intakes ($\mathrm{Max}\left(\mathrm{W}_{\mathrm{A_N}}\right) = 845.75$). Non-consumers represent around 40% of subjects in self-reported intake data, while zero value biomarker-derived intakes are only 17% of the total.

Concordance between self-reported and biomarker-derived intakes was confirmed by the statistics summarizing the conditional distributions $\mathrm{M}_{\mathrm{A_n}|\mathrm{W}_{\mathrm{A_n}}=0}$ and $\mathrm{W}_{\mathrm{A_n}|\mathrm{M}_{\mathrm{A_n}}=0}$ in

Table *1*. These distributions represent, respectively, biomarker-derived intake values corresponding to null self-reported intake values and vice-versa. Such measures of intake display good agreement, as both take low values when the other is null.

Summaries of the predicted intakes obtained using different model specifications and biomarker transformations are compared in

Table 2. Models using a log-transform of the biomarker-derived estimates of intake in the estimation of the calibration equations, or of a Box-Cox transformation with small λ parameter, result in high-variability in the calibrated intakes. Indeed, regardless of the functional specification of the calibration equations, in these cases the standard deviation of the calibrated intakes is always greater than 792, much larger than the standard deviations of the observed self-reported intakes ($\mathrm{sd}\left(\mathrm{W}_{\mathrm{A}_{\mathrm{n}^*}}\right) = 107$, see Table *1*). Also, when using these biomarker transformations, poor agreement is found between the interquartile ranges of the observed self-reported intakes and the calibrated intakes. The calibrated intakes from the remaining model specifications and biomarker transformations range in similar intervals, closer to that of the self-reported mean daily citrus intakes. Among these, the calibrated intakes computed using a linear regression model for the calibration equations with original-scale biomarker-derived estimates of intake (M_i) are those whose mean is closer to the mean of $\mathrm{W}_{\mathrm{A}_{\mathrm{n}^*}}$. Indeed, a linear regression model was optimal, in terms of MSE, to estimate calibration equations of citrus intake ($R^2 = 0.22$).

Table *3* details the average MSE values associated with different model specifications and transformations of biomarker-derived estimates of intake. Mean squared errors have been computed between biomarker-derived estimates of intakes and calibrated mean-daily self-reported intakes, computed with different specifications of the calibration equations. The lowest MSE (14354) corresponds to a linear regression model, defined with biomarker-derived estimates of intakes on the original scale, $\mathrm{M_i}$. The preference for a simple model is confirmed by the second lowest MSE, associated with a linear regression model, with unitary parameter Box-Cox transformation of $\mathrm{M_i}$, coinciding with a mere translation of the data. Given the selected optimal linear model, mean-daily calibrated citrus intakes were computed using the estimated calibration equations:

$$\hat{X}_i = 33.60 + 0.63W_i$$

It is worth noting that the estimated intercept is close to the average biomarker-derived intake for non-consumers (see Table *1*). **Figure 1** offers a comparison between calibrated and self-reported mean-daily citrus intakes for the n^* subjects in the validation study. Overall, there is good agreement between self-reported and calibrated mean-daily intakes, however, differences emerge at high intakes. Self-reported non-consumers were assigned to a low baseline citrus consumption level, corresponding to $\hat{\beta}_0 = 33.60$.

What portion of biomarker data is required in large population studies?

Sixteen distinct scenarios of simulated intake data were generated by considering different combinations of variability of self-reported data and biomarker data around the true intake, quantified through $\alpha_{\mathrm{W}}, \alpha_{\mathrm{M}} = (0.5,2)$, and self-reported data quality levels, quantified through $\beta_{\mathrm{W}} = (0.1,0.5,0.8,1)$. Among these, the group of “best-case” scenarios corresponds

to those with a $\alpha_W = \alpha_M = 0.5$, as both biomarker-derived intakes and self-reported intakes have very little variability around true intakes.

Figure *2* reports the standard deviations of the estimated calibration equations slope parameter $\hat{\beta}_1$ under such scenarios, for the different self-reported data quality levels. Standard deviations stabilize quickly, for relatively small values of n_g (the number of subjects for which biomarker-derived estimates of intake are available), regardless of the β_W value. However, β_W does influence the level at which the standard deviation stabilizes, with poor-quality self-reported data ($\beta_W = 0.1$) leading to slightly higher variability in the estimate of β_1. Similar results are found also for the set of "worst-case" scenarios (**Figure 3**), represented by the $\alpha_W = \alpha_M = 2$ configuration, that is, when large measurement error is present in both measures of intake. Standard deviations tend to stabilize quite rapidly around $n_g \approx 25000$, as happened in the "best-case" scenarios. However, this time the quality of self-reported data does not have a large impact. Indeed, there is little visible differences between the cases $\beta_W = 0.1$ and $\beta_W = 1$. The remainder of the sixteen scenarios represent less extreme, hybrid situations in terms of data quality and measurement error, with corresponding results in-line with those described for the two extreme scenarios. Similar conclusions were drawn for the estimated calibration equations intercept parameter β_0; standard deviations for the estimates of the intercept parameter stabilized jointly with those of β_1, around the same n_g values. More detail regarding the simulation study, including results and MSEs associated with the different simulation scenarios, are provided in the supplement. In general, the standard deviations of the calibration equation parameters tend to stabilize when biomarker data is obtained from 25% of the study population, in this case for 25000 out of the 100000 subjects.

DISCUSSION

The present study presents an important advancement in the use of intake biomarkers in nutrition research. Using a well-established biomarker of citrus intake calibration equations were developed to enable correction of self-reported intake for measurement error. Development and demonstration of such an approach for food intake clearly shows how one could use food intake biomarkers in nutrition research. To date the literature is lacking real examples of the utility of such food intake biomarkers with the majority of studies stopping at an association between food intake and biomarker levels. Within this context, the use of urinary proline betaine to correct self-reported citrus intake is a clear example of the potential of the field of food intake biomarkers. Furthermore, we developed a framework for assessing the amount of biomarker data needed in a large population study to correct for measurement error in food intake. Interestingly, in the present simulation study biomarker data was necessary for around 20-30% of the population highlighting an important aspect in the translation of this work into large epidemiological studies.

To date, calibration equations for correction of measurement error have focused on nutrients and to the best of our knowledge there is no example in relation to specific food intake. As dietary guidelines move towards food-based guidelines it is imperative that we also have good assessment of food intake. Use of food intake biomarkers has gained increased traction in recent years with the promise of improvement in dietary assessment. The current development of calibration equations for citrus intake clearly paves the path forward in this regard. The ability to correct for measurement error for food intake has the potential to enhance our understating of the relation between food and disease. Previous literature demonstrated that correction of measurement error in nutrient intake using biomarkers led to the identification of relationships that were not observed without the calibration. Using

biomarkers to correct for measurement error, energy intake was associated with increased risk of breast cancer, all cancer and type 2 diabetes and protein intake was associated with increased risk of type 2 diabetes (4, 18-20). Building upon the present work, the next step will be to examine relationships between calibrated food intake and certain health parameters.

An important aspect of our approach was the development of calibrated intake distributions which incorporate the uncertainty inherent in statistical modelling of measurement error. Such an approach results in calibrated intake distributions instead of single point-estimates. For researchers the potential distribution of intake delivers valuable information considering the uncertainty surrounding self-reported data. To aid the end user to implement the citrus intake calibration equations herein we developed a freely available web application, "Bio-Intake". Bio-intake is an open-source, easy to use web application designed to allow reproducibility of the methodology proposed in this paper.

Typically, for biomarker-based studies the biomarker data is acquired on a sub-population to develop the calibration equations. In this context we developed a framework for the assessment of the portion of food intake biomarker data necessary to develop accurate calibration equations. The framework advocates for examination of the quality of the self-reported data and the biomarker derived data. The framework was illustrated through a stimulation study where the portion of biomarker measures ($\mathrm{n_g}$) necessary to enable accurate development of calibration equations in a large population group ($\mathrm{N\ individuals}$) was estimated. For this purpose, standard deviations of the calibration equations parameter estimates are analyzed. A stabilization of such standard deviations occurring for low values of $\mathrm{n_g}$, $\mathrm{n_g} \ll \mathrm{N}$, would indicate that the collection of biomarker data for a small portion of the population of interest and that for the whole population are practically equivalent in terms of stability of calibration equation estimation.

Indeed, simulation results showed rapidly decreasing standard deviations, independently of the scenario considered, that is of the "observed data-quality". Such promising results suggest that substantial improvements in the correction of self-reported intake data may be achieved even with little biomarker data. Indeed, collecting biomarker-derived estimates of intake for a larger portion of the population of interest may not lead to more precise calibration equations estimates. In the current example of citrus intake, measurement of the biomarker in 20-30% of the population was deemed sufficient. The simulation study was replicated in a small cohort setting ($N = 5000$), leaving unchanged the results. Obtaining biomarker data from just a small portion of the population suggests that calibration equations can prove to be powerful tools in large cohort studies, where cost and limited biological samples are barriers to their development. In the literature, previous studies have used biomarker data in 1-3 % of the total larger cohort to assess relationships between intake and disease risk diabetes (4, 21).

The findings of the current simulation study were partly in line with previous large population studies for nutrients, in that it is not necessary to collect biomarker data for the whole population of interest to develop calibration equations. Nonetheless, simulation results showed a substantial decrease in the standard deviations of the calibration equations parameter estimates for portion of biomarker data in between $\mathrm{n_g} = 0.01\mathrm{N}$ and $\mathrm{n_g} = 0.25\mathrm{N}$, suggesting that a value of 1% may be suboptimal. Differences between subsequent standard deviations values can be used as proxies of improvement in the stability of calibration equations. Computing these differences in standard deviations for subsequent n_g values shows that they become (approximately) null for a portion of the population between 7% and 18%, depending on the scenario considered. While relaxing the recommendation to 7% of cohort biomarker data allows for greater data-collection feasibility, it induces the risk of larger

uncertainty in calibration equation estimates. Moreover, the 7% of cohort biomarker data figure is based on the strong assumption of very good quality self-reported data; small deviations from this assumption can induce even larger uncertainty in the estimates. Although being potentially onerous to achieve, the more conservative recommendation of 25% of cohort biomarker data is able to provide stable estimates independently of the self-reported data quality. Also, the simulation study was developed to analyze calibration equations estimates for food biomarkers and its results may not have direct applicability in the context of nutrient biomarkers.

Lastly, mean squared error values from the simulation study demonstrated that the quality of analyzed self-reported data is quite relevant, even with high quality biomarker data. Indeed, mean squared error values associated with "low data quality" scenarios ($\beta_W = 0.1$) were always much larger than those related to "higher data quality" ones ($\beta_W = (0.5,0.8,1)$). More importantly, mean squared error distributions in "low data quality" scenarios did not vary particularly with different variability levels in biomarker data ($\alpha_W = (0.5,2)$). In conclusion, the simulation study findings suggest that calibration equations may be stably estimated even with a relatively small portion of biomarker data, but that the quality of calibrated intake predictions is strongly dependent on that of self-reported intake data. Large biomarker data samples and excellent quality biomarkers will not be enough to get reliable estimates of intake (calibrated intake) if the observed self-reported data are not reliable.

This study has a number of noteworthy strengths. The development of calibration equations was based on reliable biomarker data and good quality food intake data (estimate from 4-day semi weighted food diaries). Furthermore, the calibration equations framework developed (and the corresponding "Bio-Intake" web app) allows estimation of calibrated intake distributions, instead of single point estimates. Hence, the proposed framework allows to

144 better account for uncertainty inherent to calibrated intakes estimation: self-reported intake
145 values are associated to a range of possible calibrated values, providing richer information on
146 the quantity of interest.
147 A number of limitations need to be acknowledged. The proposed framework only accounts
148 for self-reported intake data in the estimation of calibration equations. When available, the
149 introduction of potentially pertinent subject-specific covariates in the modelling framework
150 should be considered, as it may improve estimation of the calibrated intakes. However, as a
151 drawback, controlling for different factors with covariates would lead to the estimation of less
152 generalizable calibration equations, which would have narrower applicability. Moreover,
153 future work may consider the development of citrus calibration equations in other
154 populations, to test the generalizability of the results obtained in the present analysis. Finally,
155 the simulation study provided results concerning the stability of calibration equations
156 parameter estimates, in terms of portion of biomarker data needed for the estimation. This
157 portion does not guarantee convergence of such estimates, that is, reliable estimated
158 calibrated intakes. The results demonstrate that reliability of the calibrated intakes depends
159 on the quality of both the biomarker-derived and self-reported intake data, which does not
160 depend on the modelling framework adopted for the development of calibration equations.
161 Good self-reported data quality should always be assessed separately and before performing
162 estimating calibration equations.
163 In conclusion, the present study demonstrates the potential of food intake biomarkers in
164 nutritional epidemiology. Future work will include development of calibration equations for
165 other foods which will in turn improve the accuracy of dietary assessment. A further
166 important aspect to be explored is whether the calibrated food intakes could be used in

disease-risk studies in larger cohorts in a similar fashion to the work performed with calibrated nutrient intake.

Acknowledgements

The authors' responsibilities were as follows—SD'A: conducted research, analyzed data and prepared manuscript draft; ICG: conducted research, analyzed data and contributed to manuscript correcting; BAM, APN, JW and AF: provided essential data; LB: designed research, conducted research, analyzed data and contributed to manuscript correcting. All authors read and approved the final manuscript. None of the authors had a conflict of interest.

References

1. Bingham SA. Biomarkers in nutritional epidemiology. Public Health Nutr 2002;5(6a):821-7. doi: 10.1079/PHN2002368.
2. Kipnis V, Midthune D, Freedman L, Bingham S, Day NE, Riboli E, Ferrari P, Carroll RJ. Bias in dietary-report instruments and its implications for nutritional epidemiology. Public Health Nutr 2002;5(6a):915-23. doi: 10.1079/PHN2002383.
3. Dhurandhar NV, Schoeller D, Brown AW, Heymsfield SB, Thomas D, Sørensen TIA, Speakman JR, Jeansonne M, Allison DB, Energy balance measurement: when something is not better than nothing. Int J Obes 2015;39(7):1109-13. doi: 10.1038/ijo.2014.199.
4. Tinker LF, Sarto GE, Howard BV, Huang Y, Neuhouser ML, Mossavar-Rahmani Y, Beasley JM, Margolis KL, Eaton CB, Phillips LS, et al. Biomarker-calibrated dietary energy and protein intake associations with diabetes risk among postmenopausal women from the Women's Health Initiative. Am J Clin Nutr 2011;94(6):1600-6. doi: 10.3945/ajcn.111.018648.
5. Prentice RL, Huang Y, Kuller LH, Tinker LF, Horn LV, Stefanick ML, Sarto G, Ockene J, Johnson KC. Biomarker-calibrated energy and protein consumption and cardiovascular disease risk among postmenopausal women. Epidemiology 2011;22(2):170-9. doi: 10.1097/EDE.0b013e31820839bc.
6. Lampe JW, Huang Y, Neuhouser ML, Tinker LF, Song X, Schoeller DA, Kim S, Raftery D, Di C, Zheng C, et al. Dietary biomarker evaluation in a controlled feeding study in women from the Women's Health Initiative cohort. Am J Clin Nutr 2017;105(2):466-75. doi: 10.3945/ajcn.116.144840.
7. Gibbons H, Michielsen CJR, Rundle M, McNulty BA, Nugent AP, Gibney MJ, Brennan L, et al. Demonstration of the utility of biomarkers for dietary intake assessment; proline betaine as an example. Mol Nutr Food Res 2017;61(10).
8. Garcia-Perez I, Posma JM, Chambers ES, Nicholson JK, C. Mathers J, Beckmann M, Draper J, Holmes E, Frost G. An analytical pipeline for quantitative characterization of dietary intake: application to assess grape intake. J Agric Food Chem 2016;64(11):2423-31. doi: 10.1021/acs.jafc.5b05878.
9. Garcia-Perez I, Posma JM, Gibson R, Chambers ES, Hansen TH, Vestergaard H, Hansen T, Beckmann M, Pedersen O, Elliott P, et al. Objective assessment of dietary patterns by use of metabolic phenotyping: a randomised, controlled, crossover trial. Lancet Diabetes Endocrinol 2017;5(3):184-95. doi: 10.1016/S2213-8587(16)30419-3.
10. Pujos-Guillot E, Hubert J, Martin J-F, Lyan B, Quintana M, Claude S, Chabanas B, Rothwell JA, Bennetau-Pelissero C, Scalbert A, et al. Mass spectrometry-based metabolomics for the discovery of biomarkers of fruit and vegetable intake: citrus fruit as a case study. J Proteome Res 2013;12(4):1645-59. doi: 10.1021/pr300997c.
11. Heinzmann SS, Brown IJ, Chan Q, Bictash M, Dumas M-E, Kochhar S, Stamler J, Holmes E, Elliott P, Nicholson JK. Metabolic profiling strategy for discovery of nutritional biomarkers: proline betaine as a marker of citrus consumption. Am J Clin Nutr 2010;92(2):436-43. doi: 10.3945/ajcn.2010.29672.
12. Lloyd AJ, Beckmann M, Favé G, Mathers JC, Draper J. Proline betaine and its biotransformation products in fasting urine samples are potential biomarkers of habitual citrus fruit consumption. Br J Nutr 2011;106(6):812-24. doi: 10.1017/S0007114511001164.
13. Gormley IC, Bai Y, Brennan L. Combining biomarker and self-report dietary intake data: a review of the state of the art and an overview of concepts. Stat Methods Med Res (in press) 2019.
14. Zhang S, Midthune D, Guenther PM, Krebs-Smith SM, Kipnis V, Dodd KW, Buckman DW, Tooze JA, Freedman L, Carroll RJ. A new multivariate measurement error model with zero-inflated dietary data, and its application to dietary assessment. Annals Appl Stat 2011;5(2B):1456-87. doi: 10.1214/10-AOAS446.

15. Prentice RL, Mossavar-Rahmani Y, Huang Y, Van Horn L, Beresford SAA, Caan B, Tinker L, Schoeller D, Bingham S, Eaton CB, et al. Evaluation and comparison of food records, recalls, and frequencies for energy and protein assessment by using recovery biomarkers. Am J Epidemiol 2011;174(5):591-603. doi: 10.1093/aje/kwr140.
16. Kipnis V, Midthune D, Buckman DW, Dodd KW, Guenther PM, Krebs-Smith SM, Subar AF, Tooze JA, Carroll RJ, Freedman LS. Modeling data with excess zeros and measurement error: application to evaluating relationships between episodically consumed foods and health outcomes. Biometrics 2009;65(4):1003-10. doi: 10.1111/j.1541-0420.2009.01223.x.
17. Keogh RH, White IR. A toolkit for measurement error correction, with a focus on nutritional epidemiology. Stat Med 2014;33(12):2137-55. doi: 10.1002/sim.6095.
18. Prentice RL, Pettinger M, Tinker LF, Huang Y, Thomson CA, Johnson KC, Beasley J, Anderson G, Shikany JM, Chlebowski RT, et al. Regression calibration in nutritional epidemiology: example of fat density and total energy in relationship to postmenopausal breast cancer. Am J Epidemiol 2013;178(11):1663-72. doi: 10.1093/aje/kwt198.
19. Prentice RL, Shaw PA, Bingham SA, Beresford SAA, Caan B, Neuhouser ML, Patterson RE, Stefanick ML, Satterfield S, Thomson CA, et al. Biomarker-calibrated energy and protein consumption and increased cancer risk among postmenopausal women. Am J Epidemiol 2009;169(8):977-89. doi: 10.1093/aje/kwp008.
20. Zheng C, Beresford SA, Van Horn L, Tinker LF, Thomson CA, Neuhouser ML, Di C, Manson JE, Mossavar-Rahmani Y, Seguin R, et al. Simultaneous association of total energy consumption and activity-related energy expenditure with risks of cardiovascular disease, cancer, and diabetes among postmenopausal women. Am J Epidemiol 2014;180(5):526-35. doi: 10.1093/aje/kwu152.
21. Neuhouser ML, Tinker L, Shaw PA, Schoeller D, Bingham SA, Horn LV, Beresford SAA, Caan B, Thomson C, Satterfield S, et al. Use of recovery biomarkers to calibrate nutrient consumption self-reports in the Women's Health Initiative. Am J Epidemiol 2008;167(10):1247-59. doi: 10.1093/aje/kwn026.

Table 1

CITRUS INTAKE	W_{A_N}	$W_{A_{n^*}}$	W_{A_n}	M_{A_n}	$M_{A_n}\|_{W_{A_n}=0}$	$W_{A_n}\|_{M_{A_n}=0}$
Mean	75.42	72.41	80.39	83.26	30.52	24.18
Standard deviation	106.11	107.19	104.21	135.45	56.56	54.01
Min	0.00	0.00	0.00	0.00	0.00	0.00
First quartile	0.00	0.00	0.00	7.83	0.00	0.00
Median	31.88	22.00	43.75	34.48	12.81	0.00
Third quartile	117.31	113.00	122.75	98.86	35.07	7.12
Max	845.75	845.75	712.50	1226.79	441.20	302.50
Zero intakes	646 (43%)	424 (45%)	215 (38%)	94 (17%)	68 (32%)	68 (72%)

Summary statistics for the citrus mean daily intake data (g/day). Self-reported mean daily intakes are reported for the entire sample (W_{A_N}), subjects for which no biomarker data is available ($W_{A_{n^*}}$) and subjects for which biomarker-derived estimates of intake are computed (W_{A_n}). The fourth column reports summary statistics for the biomarker-derived estimates of intake (M_{A_n}).

Table 2

	MODEL	Linear regression model							Tobit model		Two-part model	
	BIOMARKER TRANSFORMATION	M_i	$log(M_i)$	$BC(M_i, 0.1)$	$BC(M_i, 0.5)$	$BC(M_i, 1)$	$BC(M_i, 1.1)$	$BC(M_i, 1.5)$	M_i	$log(M_i)$	M_i	$log(M_i)$
CALIBRATED MEAN DAILY CITRUS INTAKE (g/day)	Mean	81	103	116	58	81	86	101	81	66	64	128
	Standard deviation	66	792	1299	69	66	69	82	67	180	77	1332
	Min	34	19	3	22	34	35	33	33	17	8	15
	First quartile	34	19	3	22	34	35	33	33	17	8	15
	Median	54	25	4	31	53	58	74	52	25	32	20
	Third quartile	107	52	19	66	107	115	150	108	62	94	45
	Max	563	24124	38554	891	563	549	527	553	4313	625	42066

Summary statistics for calibrated intake values using different calibration equations specifications, corresponding to various biomarker-derived intake data (M_i) transformations and functional specifications. The calibrated intakes are those computed for the group of n^*subjects for which only self-reported mean daily citrus intake data (g/day) were available.

Table 3

Average Mean Square Error			
	BIOMARKER TRANSFORMATION		
MODEL	$\mathbf{M_i}$	$\mathbf{\log(M_i + 1)}$	$\mathbf{BC(M_i, 1)}$
Linear regression model	14354	49078	14480
Tobit model	15034	89645	-
Two-part model	21504	21010	-

Average mean squared errors computed between calibrated intakes and biomarker-derived estimates of intakes (M_i), for the n subjects in the study and different specifications of the calibration equations. The average MSE values for different specification of the Box-Cox λ parameter have been omitted, being all greater than the one obtained with $\lambda = 1$.

Figure Legends

Figure 1 Self-reported mean daily citrus intakes (g/day) and calibrated mean daily citrus intakes (g/day). For the calibrated intakes, 95% confidence intervals are reported (light grey). Intakes are ordered from lowest to highest values.

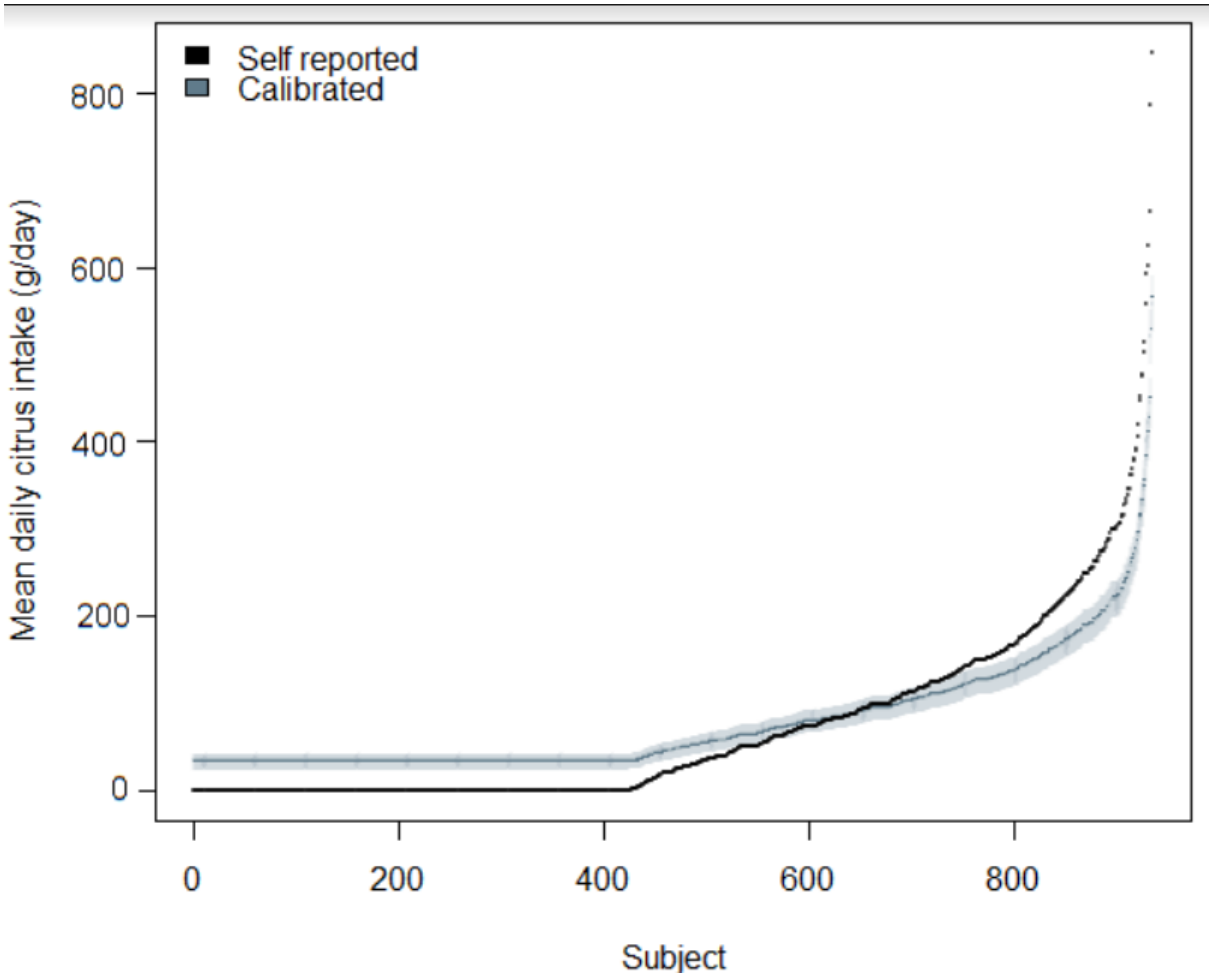


Figure 2 Standard deviations of the estimated slope coefficient β_1 are reported, for different sample sizes n_g and different values of the β_W parameter. These simulated scenarios represent the "best possible" ones in terms of data quality, as both self-reported and biomarker-derived estimates of intakes have low variability around the true intake values. Notice that the standard deviations tend to stabilize quickly, even in the "worst-case" scenario, that is when $\beta_W = 0.1$.

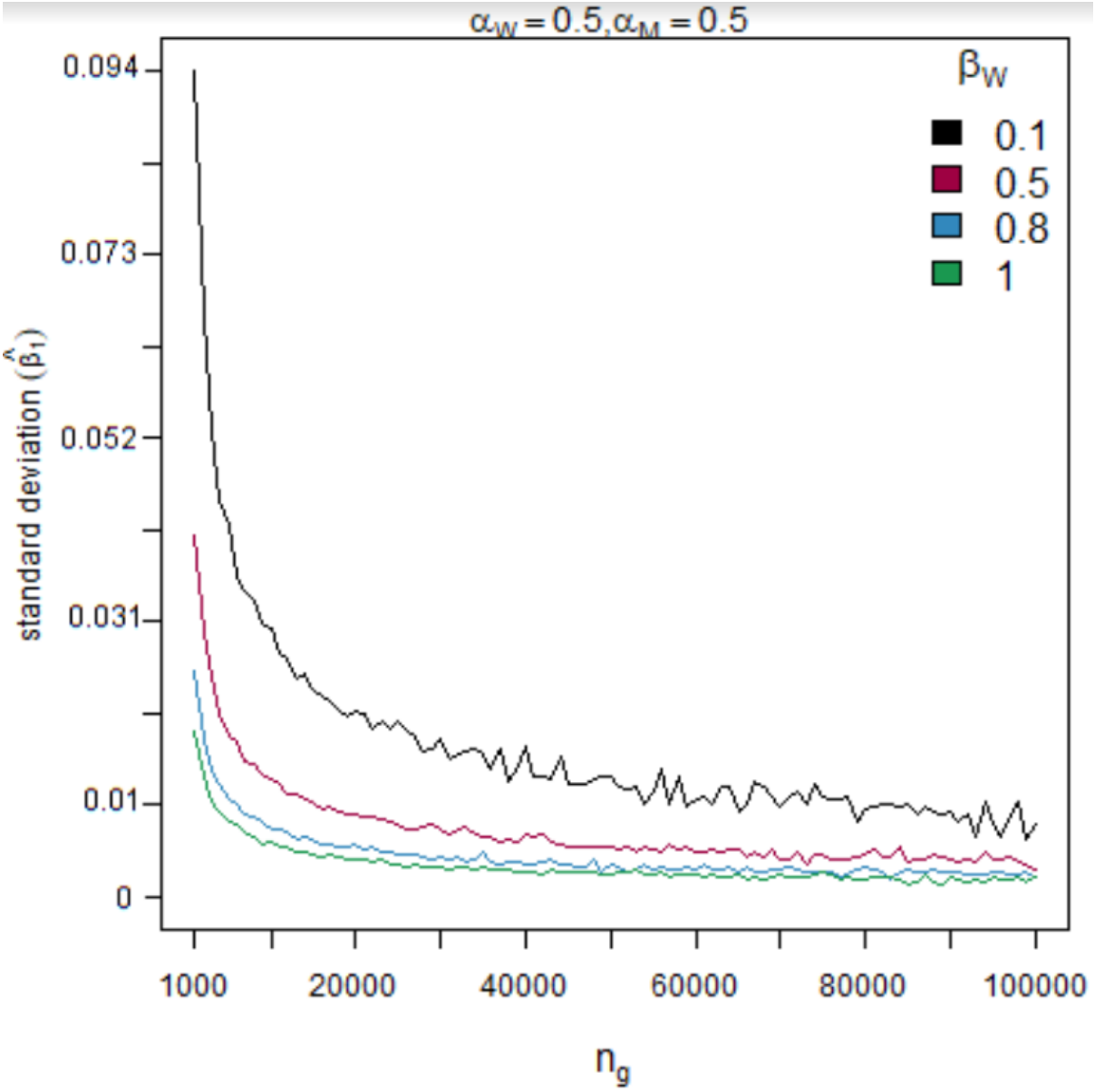
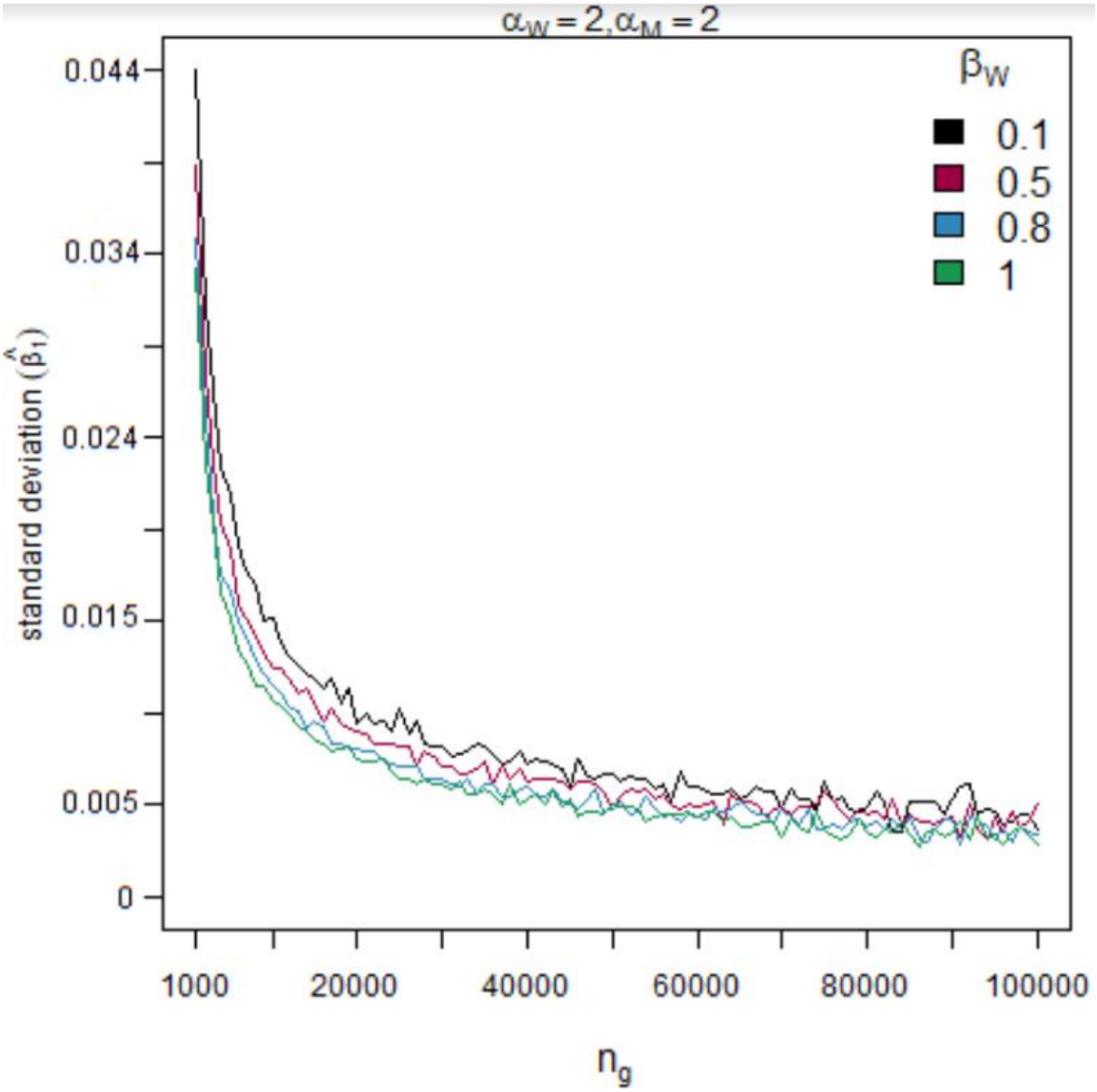


Figure 3 Standard deviations of the estimated slope coefficient β_1 are reported, for different sample sizes n_g and different values of the β_W parameter. These simulated scenarios represent the "worst possible" outcomes in terms of data quality, as both self-reported and biomarker-derived estimates of intakes have high variability around the true intake values. Notice that the standard deviations tend to stabilize quickly, even in the "worst-case" scenario, that is when $\beta_W = 0.1$.


$\alpha_W = 2, \alpha_M = 2$
β_W
0.1
0.5
0.8
1
standard deviation ($\hat{\beta}_1$)
0.044
0.034
0.024
0.015
0.005
0
1000
20000
40000
60000
80000
100000
n_g

SUPPLEMENTARY MATERIAL

COMBINING BIOMARKER AND FOOD INTAKE DATA: CALIBRATION EQUATIONS FOR CITRUS INTAKE

Silvia D'Angelo, Isobel Claire Gormley, Breige A McNulty, Anne P Nugent, Janette Walton, Albert Flynn, and Lorraine Brennan

THE CALIBRATION METHOD

The selection process for the data used to develop calibration equations is illustrated by **Supplementary Figure 4**.

How it works

The calibration method is a measurement error correction technique widely used in nutrition literature (1-3). The main assumption behind it states that self-reported intake data $\mathrm{W_i}$ are linearly related to true intakes $\mathrm{X_i}$:

$$W_i = \mu_w + \beta_w X_i + \epsilon_{W,i} \quad (1)$$

for subjects $\mathrm{i} = 1, \dots, \mathrm{N}$. The relation between the observed data $\mathrm{W_i}$ and the latent quantity of interest $\mathrm{X_i}$ is mainly influenced by two factors: the slope coefficient β_W and the subject-specific errors $\epsilon_{\mathrm{w,i}}$. The β_W parameter may be seen as a measure of the "quality" or the intensity of the relationship between self-reported and true intake data. Also, it expresses the direction of the relation between the two variables, with negative values of β_W indicating disagreement. Subject-specific errors $\epsilon_{\mathrm{w,i}}$ represent individual variability in self-reported intakes. Such errors are assumed to be normally distributed with zero mean and homoscedastic, i.e. with equal variances σ^2_w. Thus, the calibration method assumes there is no interaction between the subjects when reporting the intakes, a reasonable assumption in many nutrition studies, as reports are collected separately. Also, by assuming equal error variances σ^2_w, the calibration method implies that subjects are equally likely to misreport their consumption, hence all subjects are equally reliable.

The linear relation between self-reported and true intake data cannot be estimated, as the latter is not observed. To overcome such issue, the calibration method relies on a second assumption: another linear relation exists between biomarker-derived ($\mathrm{M_i}$) and true intake data:

$$M_i = \mu_M + \beta_M X_i + \epsilon_{M,i} \quad (2)$$

Further, it is assumed that biomarker-derived estimates of intake provide a more reliable representation of true intakes, setting $\beta_\mathrm{M} = 1$ and $\mu_\mathrm{M} = 0$, thus following the classical measurement error model REF here Carroll book 2006. Subject-specific errors $\epsilon_{\mathrm{M,i}}$ are assumed to be independently distributed, according to a Gaussian distribution with zero mean and equal variance σ^2_M.

As the true intakes $\mathrm{X_i}$ cannot be observed, the calibration method combines equations (1) and (2) to derive estimates of such quantities. Specifically, given the assumptions outlined hold, $\mathrm{X_i}$ may be conditionally predicted given biomarker-derived estimates of intakes and self-reported intakes in two steps:

$$\mathrm{M_i} = \beta_0 + \beta_1 \mathrm{W_i} + \epsilon_\mathrm{i}, \quad \mathrm{i} = 1, \dots, \mathrm{n}$$
$$\widehat{\mathrm{X_i}} = \widehat{\beta_0} + \widehat{\beta_1} \mathrm{W_i}, \quad \mathrm{i} = 1, \dots, \mathrm{n}$$

The errors ϵ_i are within-subject random effects, distributed independently according to a standard Gaussian distribution. The above equations are the so-called "calibration equations" and the predicted true intakes $\widehat{X}_i$ are the calibrated self-reported intakes. Indeed, the estimates of the fixed effect population level parameters $\widehat{\beta_0}$ and $\widehat{\beta_1}$ may be used to correct self-reported data and get more reliable estimates of true intake; note that such correction incorporates the information coming from biomarker-derived data, but it does not directly depend on it. (4) provides detail on the construction, implementation and theoretical justification of the calibration method.

Alternative model formulations

Intake data is often characterized by a large presence of zero values. Depending on whether the measure of intake is self-reported or biomarker-derived, zero values either indicate non-consumers (subjects that reported null consumption of a given food) or null biomarker-derived estimates. There may be a partial mismatch between subjects self-reporting zero intake values and those having zero biomarker-derived estimates of intake values. When subject-specific covariates have been collected, several approaches, eg. (5), have been proposed to incorporate them in the modelling framework and to try modelling "true" and "false" (those misreporting) non-consumers. In the absence of covariate data such zero self-reported intakes are more challenging to address, but modelling frameworks designed for zero-inflated data can be exploited eg. (6). In particular, the calibration equations may be re-defined according to either a Tobit model (7) or a two-part model (8); such models are designed specifically for a regression framework where the response variable, M_i, is semi-continuous. Then, the calibration equations may be rewritten as:

$$M_i = f_l(W_i) + \epsilon_i, \quad i = 1, \dots, n$$
$$\widehat{X}_i = \widehat{f_l}(W_i)$$

The function $f_l(\cdot)$ is a link function, corresponding to the linear regression, the Tobit regression or the two-part regression, respectively for $l = 1,2,3$.

Biomarker transformation

The assumption of Gaussian errors ϵ_i may not be completely satisfied in empirical studies, and some skewness may be found in observed intake data. As can be seen from the summary statistics provided in the paper, this is the case for the observed mean daily self-reported intakes and biomarker-derived estimates of citrus intake. In particular, when deriving calibration equations for citrus intake, the skewness in the distribution of biomarker-derived estimates of intake may imply non-Normality of the errors ϵ_i. To try overcome this potential issue, in the context of episodically consumed foods (5) propose to transform the response variable using the Box-Cox family of transformations (9),

$$M_i^* = BC(M_i, \lambda) = \frac{M_i^\lambda - 1}{\lambda}, \quad i = 1, \dots, n$$

so that the distribution of the transformed marker data M_i^* would more closely resemble that of a Gaussian. Using a Box-Cox transformation of biomarker-derived intakes, the calibration equations may be defined as:

$$M_i^* = f_l(W_i) + \epsilon_i, \quad i = 1, \dots, n$$
$$\widehat{X}_i^* = \widehat{f_l}(W_i)$$

where $\widehat{X}_i^*$ is the transformed predicted true intake and $f_l(\cdot)$ is any given model specification for the calibration equations. Predicted intakes on the original scale may be found by computing the inverse of the Box-Cox transformation, $\widehat{X_i} = BC^{-1}(\widehat{X}_i^*, \lambda)$, using the Delta method (10):

$$\widehat{\mathrm{X}}_{\mathrm{i}} = \left(\lambda\widehat{\mathrm{X}}_{\mathrm{i}}^{*} + 1\right)^{\frac{1}{\lambda}}\left[1 + \frac{(1-\lambda)}{2\left(\lambda\widehat{\mathrm{X}}_{\mathrm{i}}^{*} + 1\right)^{2}}\right]$$

Note that when biomarker-derived intake data are transformed using a Box-Cox transformation, the two-part and Tobit model can no longer be estimated, as structural zeros "disappear" in the transformed variable.

A different transformation that attempts to normalize the data while leaving the structural zeros unchanged is the log-transformation $\mathrm{M}_{\mathrm{i}}^{*} = \log(\mathrm{M}_{\mathrm{i}} + 1)$. This type of log-transformation is widely used in dose-response study data, where there is often the need to rescale the data without losing information on the structural zero values. Indeed, for $\mathrm{M}_{\mathrm{i}} = 0$, $\mathrm{M}_{\mathrm{i}}^{*} = \log(1) = 0$. As for Box-Cox transformed intake data, when using log-transformed biomarker-derived estimates of intake to estimate calibration equations, the computed calibrated intakes are defined on a new, transformed, scale. Calibrated intake on the original scale may be obtained using the back transformation $\widehat{\mathrm{X}}_{\mathrm{i}} = \frac{3}{2}\exp\left(\widehat{\mathrm{X}}_{\mathrm{i}}^{*}\right) - 1$.

Estimating calibration equations: cross validation and mean squared error

In order to select the optimal model specification, $\hat{\mathrm{f}}_{\mathrm{l}}(\cdot)$, $\mathrm{l} = 1,2,3$, and data transformation (original scale, Box-Cox or log-transformation), different calibration equation models have been estimated, resulting from all possible combinations of model specifications and data transformations. For the Box-Cox transformation, parameter values $\lambda = 0.05, 0.1, 0.15, \dots, 1.2$ have been considered, as these yield the most Gaussian-like distributions of transformed biomarker derived estimates of intake. As a result, 30 different models have been estimated. To obtain stable estimates of the calibration equations, each model $\mathcal{M}_{\mathrm{k}}$, $\mathrm{k} = 1, \dots, \mathrm{K}$ ($\mathrm{K} = 30$), has been estimated using "leave-p-out cross-validation", with "p" corresponding to 20% of the data. Such operation has been repeated $\mathrm{T} = 5000$ times, $\mathrm{t} = 1, \dots, \mathrm{T}$, each time randomly splitting the observations in A_{n} into a training set (containing 80% of the data) and a test set. For each $\mathcal{M}_{\mathrm{k}}$ model, calibration equations have been estimated using training data only; the corresponding estimated coefficients are then used to compute calibrated intakes from the self-reported intakes in the test set:

$$M_{i,k,t}^{train} = f_l\left(W_{i,k,t}^{train}\right) + \epsilon_{i,k,t}^{train} \tag{3}$$

$$\hat{X}_{i,k,t}^{test} = \hat{f}\left(W_{i,k,t}^{test}\right) \tag{4}$$

When using transformed biomarker-derived estimates of intake, in equations (3) and (4), $\mathrm{M}_{\mathrm{i,k,t}}^{\mathrm{train}}$ may be substituted with the transformed-intake $\mathrm{M}_{\mathrm{i,k,t}}^{\mathrm{train},*}$, and $\widehat{\mathrm{X}}_{\mathrm{i,k,t}}^{\mathrm{test}}$ would then be the transformed-calibrated intake $\widehat{\mathrm{X}}_{\mathrm{i,k,t}}^{\mathrm{test},*}$.

To compare calibrated intakes obtained with different calibration equations models $\mathcal{M}_{\mathrm{k}}$, $\mathrm{k} = 1, \dots, \mathrm{K}$, mean squared errors are computed, between the known biomarker-derived intake values $\mathrm{M}_{\mathrm{i,k,t}}^{\mathrm{test}}$ and the calibrated intake values $\widehat{\mathrm{X}}_{\mathrm{i,k,t}}^{\mathrm{test}}$:

$$\mathrm{MSE}_{\mathrm{k,t}} = \frac{\sum_{\mathrm{i}=1}^{0.20\mathrm{n}}\left(\mathrm{M}_{\mathrm{i,k,t}}^{\mathrm{test}} - \widehat{\mathrm{X}}_{\mathrm{i,k,t}}^{\mathrm{test}}\right)^{2}}{0.20\mathrm{n}}$$

Different models may then be compared using average mean square errors $\overline{\mathrm{MSE}_{\mathrm{k}}} = \sum_{\mathrm{t}=1}^{\mathrm{T}}\mathrm{MSE}_{\mathrm{k,t}}/\,\mathrm{T}$, with lower values corresponding to better models. Thus, the optimal model $\mathcal{M}_{\mathrm{k}^{*}}$ may be found as $\mathrm{k}^{*} = \arg\min_{\mathrm{k}}\overline{\mathrm{MSE}_{\mathrm{k}}}$.

SIMULATION STUDY

The calibration equations framework is built upon the assumption that both self-reported intakes and biomarker-derived estimates of intake are linearly related to the unknown true intakes. Normally, biomarker-derived intakes are collected only for a sub-sample of the population of interest, due to elevated extraction costs. Indeed, in the mean daily citrus intake data the population of interest counted $\mathrm{N} = 1500$ subjects, but biomarker-derived intakes had been collected only for $\mathrm{n} = 565$ of them, approximately one third of the total. With large population studies, it may be infeasible to collect biomarker-derived intakes for a large subset of the subjects. Biomarker-derived data would be then collected only for a small number of n subjects, $\mathrm{n} \ll \mathrm{N}$. However, deriving calibration equations using $\mathrm{n} \ll \mathrm{N}$ observations may result in unstable and potentially unreliable intake predictions. Hence, how large, or small, n is required to be in order to obtain stable predictions is of key interest. To investigate this issue, a simulation study has been constructed. Assuming the population of interest has $\mathrm{N} = 100{,}000$ subjects, $\mathrm{P} = 100$ different sets of true, self-reported and biomarker-derived mean daily citrus intakes have been simulated according to the calibration equations framework:

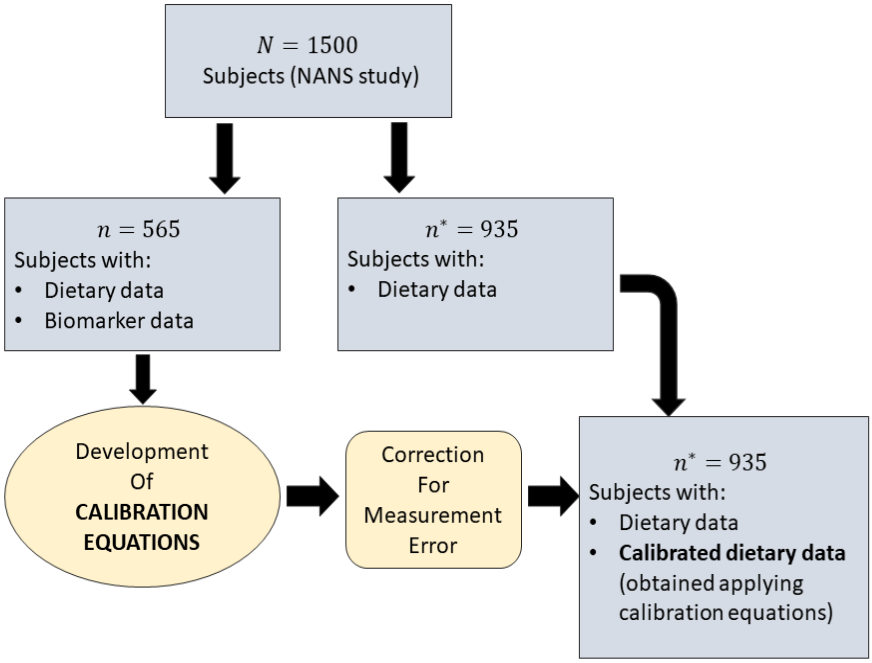
$$\mathrm{W}_{\mathrm{i,p}} = \mu_{\mathrm{W}} + \beta_{\mathrm{W}}\mathrm{X}_{\mathrm{i,p}} + \epsilon_{\mathrm{i,p,W}}$$
$$\mathrm{M}_{\mathrm{i,p}} = \mathrm{X}_{\mathrm{i,p}} + \epsilon_{\mathrm{i,p,M}}$$

True intakes $\mathrm{X}_{\mathrm{i,p}}$ have been simulated from a Gaussian distribution $\mathrm{X}_{\mathrm{i,p}} \sim \mathcal{N}(\mu_{\mathrm{X}}, \sigma_{\mathrm{X}}^2)$, where $\mu_{\mathrm{X}} = 81$ and $\sigma_{\mathrm{X}}^2 = (66)^2$ have been fixed to reflect the estimated mean and variance of the calibrated mean daily citrus intakes. The random errors, $\epsilon_{\mathrm{i,p,W}}$ and $\epsilon_{\mathrm{i,p,M}}$ are assumed to be normally distributed with zero mean and variances $\sigma_{\mathrm{W}}^2 = \alpha_{\mathrm{W}}\sigma_{\mathrm{X}}^2$ and $\sigma_{\mathrm{M}}^2 = \alpha_{\mathrm{M}}\sigma_{\mathrm{X}}^2$. The α_{W} and α_{M} scale parameters regulate the variation of the self-reported and biomarker-derived intake data around the true intake. These scale parameters have been set to $\alpha_{\mathrm{W}}, \alpha_{\mathrm{M}} = (0.5{,}2)$, to represent both clean ($\alpha_{\mathrm{W}}, \alpha_{\mathrm{M}} = 0.5$) and noisy ($\alpha_{\mathrm{W}}, \alpha_{\mathrm{M}} = 2$) intake data. The β_{W} parameter measures the strength of the relation between self-reported and true intake data. To represent different levels of quality for self-reported data, this parameter has been set to $\beta_{\mathrm{W}} = (0.1{,}0.5{,}0.8{,}1)$, with $\beta_{\mathrm{W}} = 0.1$ and $\beta_{\mathrm{W}} = 1$ denoting poor and good data quality, respectively. Finally, the offset parameter μ_{W} measures the average difference between self-reported and true intake data. Here this parameter has been set to $\mu_{\mathrm{W}} = 0$, as similar values were observed for the average self-reported and calibrated mean-daily intakes.

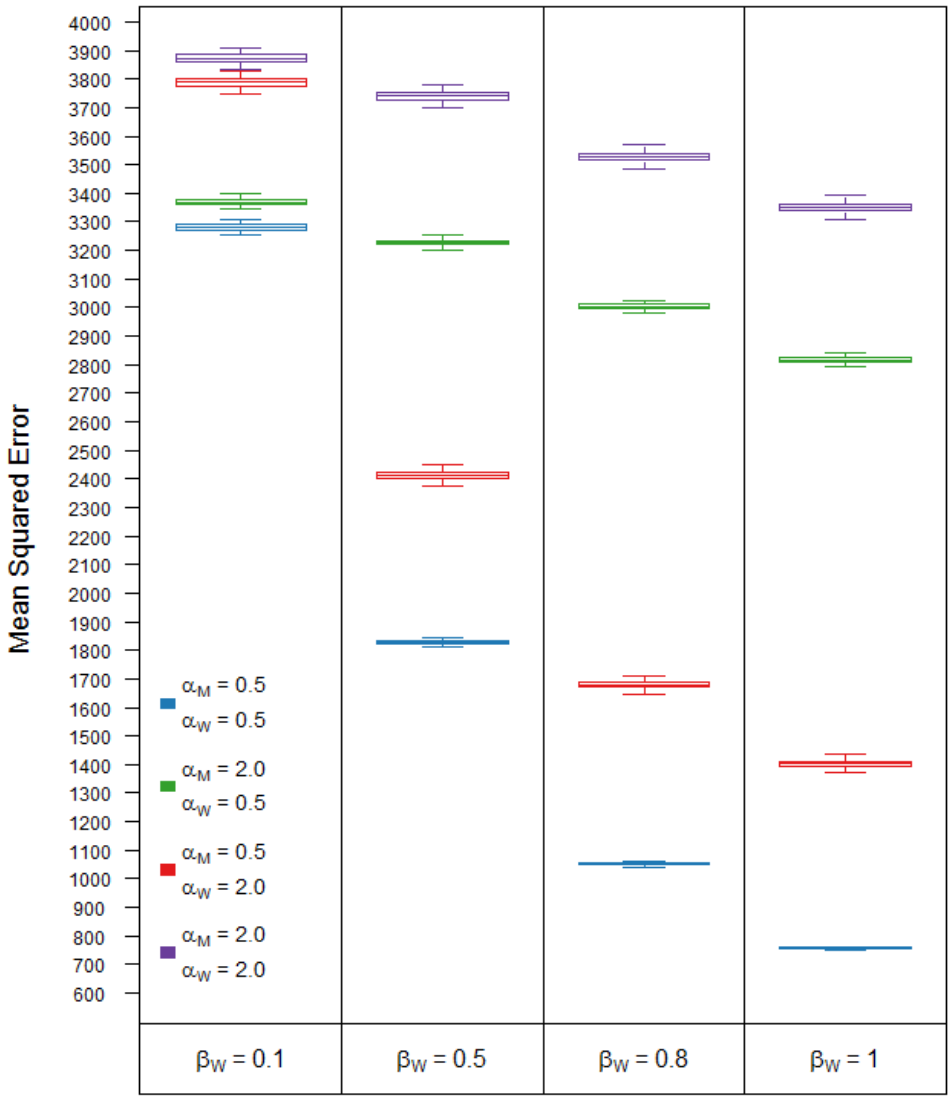
Different combinations of β_{W}, α_{W} and α_{M} parameter values give rise to sixteen different simulation scenarios. For each scenario, it has been further assumed that biomarker-derived estimates of intake have been collected for a subsample of $\mathrm{n_g}$ subjects, with $\mathrm{n_g} = \mathrm{gN}$ and $\mathrm{g} = (0.01, 0.02, \ldots, 0.99{,}1)$. For each simulation scenario and for each value of $\mathrm{n_g}$ the calibration equations are estimated, using the optimal model from the citrus intake data analysis i.e. the linear model with the original-scale biomarker-derived intakes. As in the citrus intake analysis, calibration equations have been estimated using the "leave-p-out cross-validation" method. The standard deviations of the cross-validation estimates of the calibration equation parameters, $\mathrm{sd}\left(\hat{\beta}_0\right)_{\alpha_{\mathrm{W}},\alpha_{\mathrm{M}},\beta_{\mathrm{W}},\mathrm{n_g},\mathrm{p}}$ and $\mathrm{sd}\left(\hat{\beta}_1\right)_{\alpha_{\mathrm{W}},\alpha_{\mathrm{M}},\beta_{\mathrm{W}},\mathrm{n_g},\mathrm{p}}$ are then computed. Finally, these standard deviations are averaged across the P different simulated datasets giving $\mathrm{sd}\left(\hat{\beta}_0\right)_{\alpha_{\mathrm{W}},\alpha_{\mathrm{M}},\beta_{\mathrm{W}},\mathrm{n_g}}$ and $\mathrm{sd}\left(\hat{\beta}_1\right)_{\alpha_{\mathrm{W}},\alpha_{\mathrm{M}},\beta_{\mathrm{W}},\mathrm{n_g}}$. For each scenario, these averaged standard deviations are explored for increasing number of biomarker-derived intake estimates $\mathrm{n_g}$. Small decrements of the standard deviations with increasing $\mathrm{n_g}$ would indicate no improvement in parameter estimate stability as a result of, expensively, collecting greater portions of biomarker data.

Simulation study results

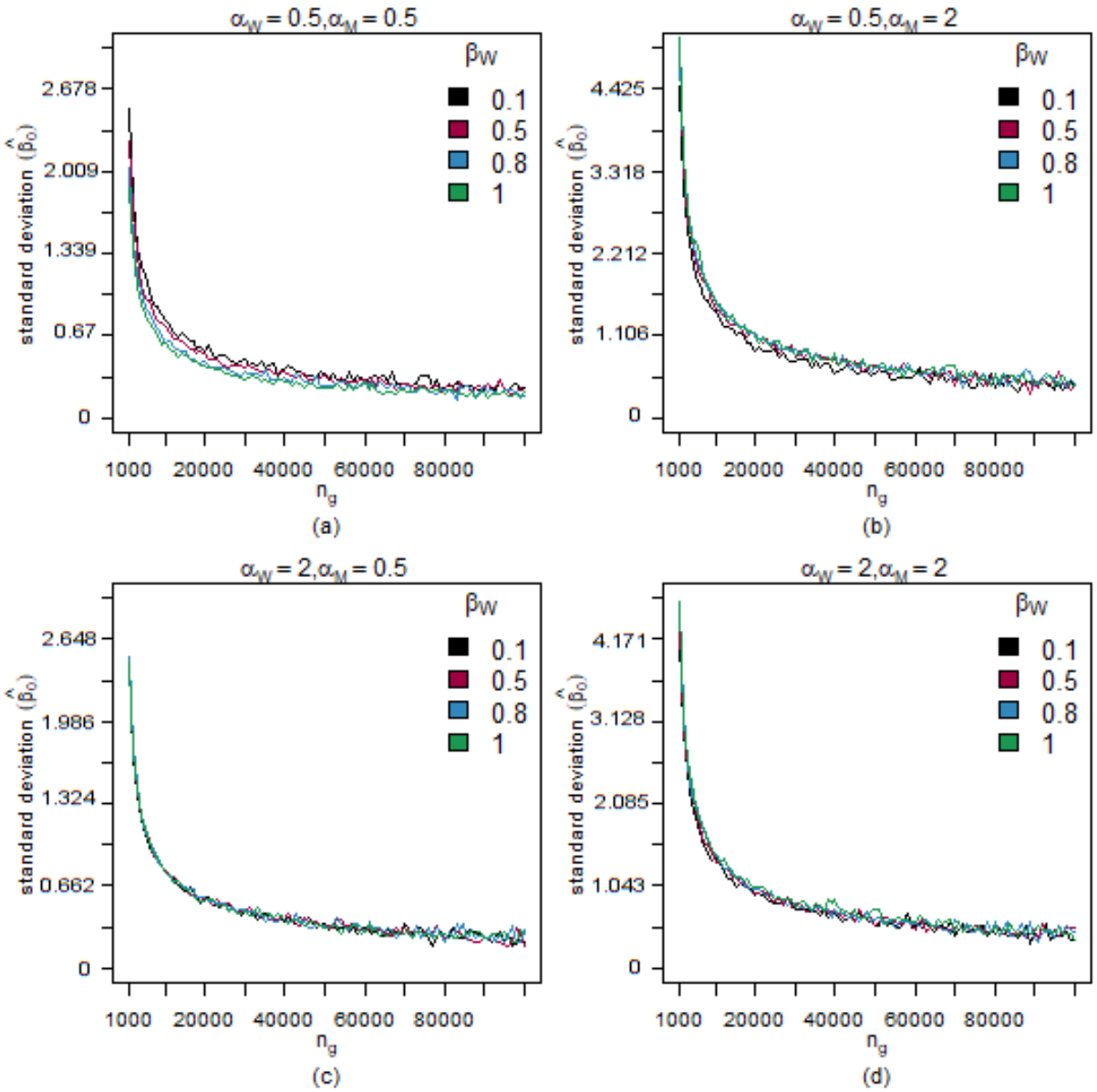
Supplementary Figure 5 illustrates the distributions of the Mean Squared Errors computed in the different simulation scenarios, between simulated and calibrated mean daily intakes. The MSE increases with decreasing data quality and with increasing variability of the self-reported and biomarker-derived intakes around the true intakes. Notice that when $\beta_W = 0.1$, the MSE is not much affected by the noise level (α_W, α_M), suggesting that the quality of self-reported data is quite relevant. **Supplementary Figure 6** reports the standard deviations of the estimated intercept β_0 in the different scenarios, while **Supplementary Figure 7** those of the estimated slope coefficient β_1 in the intermediate scenarios. Standard deviations tend to stabilize quickly, independently of the scenario considered.



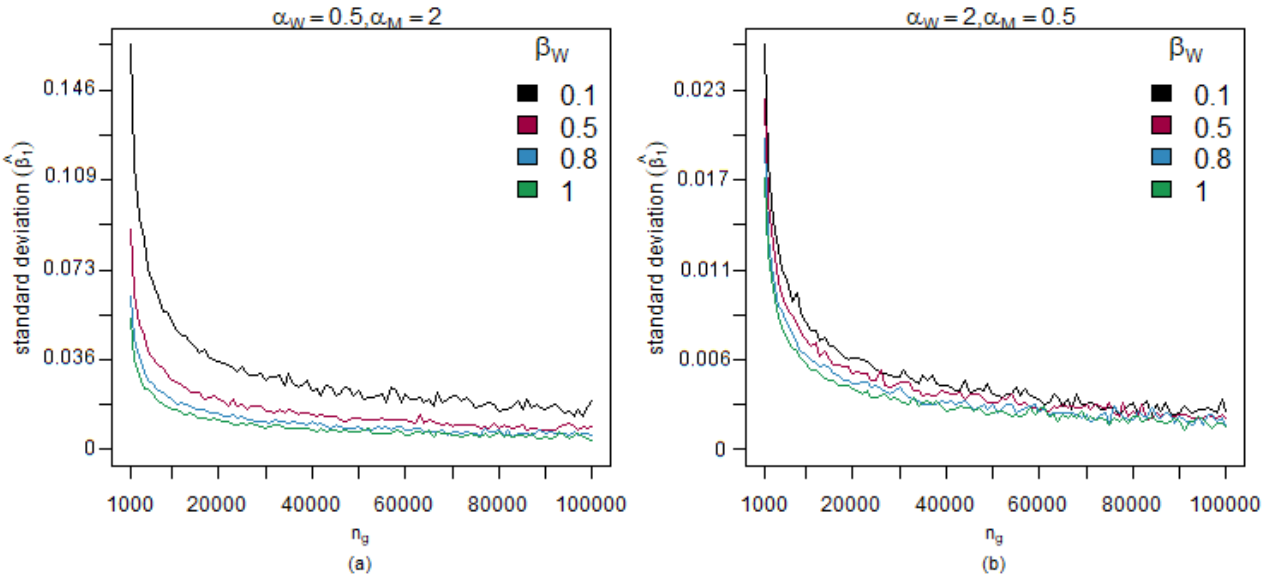
Supplementary Figure 4 Flow chart describing the selection process for the data used to develop calibration equations.



Supplementary Figure 5 Simulations. This figure presents the distributions of the MSE computed in the different simulation scenarios, between simulated and calibrated mean daily intakes. As expected, the error is lower when both self-reported and biomarker-derived intakes have little variability around true intakes, and it decreases with increasing values of β_W. However, when self-reported intake data is of "poor" quality ($\beta_W = 0.1$), the MSE is not much affected by the noise level (α_W, α_M).



Supplementary Figure 6 Standard deviations of the estimated intercept β_0 in the different scenarios, for different sample sizes n_g and different values of the β_W parameter. As for the estimated slope coefficient β_1, standard deviations tend to stabilize quickly, independently of the scenario considered.



Supplementary Figure 7 Standard deviations of the estimated slope coefficient β_1 in the two intermediate scenarios, for different sample sizes n_g and different values of the β_W parameter. As for the other scenarios, standard deviations tend to stabilize quickly, even with "poor" self-reported data quality, that is for $\beta_W = 0.1$.

References

1. Prentice RL, Pettinger M, Tinker LF, Huang Y, Thomson CA, Johnson KC, Beasley J, Anderson G, Shikany JM, Chlebowski RT, et al. Regression calibration in nutritional epidemiology: example of fat density and total energy in relationship to postmenopausal breast cancer. Am J Epidemiol 2013;178(11):1663-72. doi: 10.1093/aje/kwt198.
2. Prentice RL, Huang Y, Kuller LH, Tinker LF, Horn LV, Stefanick ML, Sarto G, Ockene J, Johnson KC. Biomarker-calibrated energy and protein consumption and cardiovascular disease risk among postmenopausal women. Epidemiology 2011;22(2):170-9. doi: 10.1097/EDE.0b013e31820839bc.
3. Huang Y, Van Horn L, Tinker LF, Neuhouser ML, Carbone L, Mossavar-Rahmani Y, Thomas F, Prentice RL. Measurement error corrected sodium and potassium intake estimation using 24-hour urinary excretion. Hypertension 2014;63(2):238-44. doi: 10.1161/HYPERTENSIONAHA.113.02218.
4. Gormley IC, Bai Y, Brennan L. Combining biomarker and self-report dietary intake data: a review of the state of the art and an overview of concepts. Stat Methods Med Res (in press) 2019.

5. Kipnis V, Midthune D, Buckman DW, Dodd KW, Guenther PM, Krebs-Smith SM, Subar AF, Tooze JA, Carroll RJ, Freedman LS. Modeling data with excess zeros and measurement error: application to evaluating relationships between episodically consumed foods and health outcomes. Biometrics 2009;65(4):1003-10. doi: 10.1111/j.1541-0420.2009.01223.x.
6. Min Y, Agresti A. Modeling nonnegative data with clumping at zero: a survey. JIRSS 2002;1(1):7-33.
7. Tobin J. Estimation of relationships for limited dependent varaibles: 1. Introduction. Econometrica (pre-1986) 1958;26(1):24.
8. Duan N, Manning WG, Morris CN, Newhouse JP. A comparison of alternative models for the demand for medical care. J Bus Econ Stat 1983;1(2):115-26. doi: 10.1080/07350015.1983.10509330.
9. Box GEP, Cox DR. An analysis of transformations. J R Stat Soc Series B Stat Methodol 1964;26(2):211-43. doi: doi:10.1111/j.2517-6161.1964.tb00553.x.
10. van der Vaart A.W. The Delta-method. In Weak Convergence and Empirical Processes. New York: Springer, 1996.