Web Appendix 1

Preliminary Data Preparations

Calculation of protein density values:

Protein density is defined as the ratio (%) of energy from protein to total energy. The self-report value was calculated as 4 (kcal/g) x reported protein (g) / reported total energy (kcal). The biomarker value was calculated as 4 x urinary measure of dietary protein (g) / DLW total energy closest in time (kcal). Each urinary protein measurement was converted to a biomarker value for protein density, so individuals with repeat biomarker values for protein also had repeat biomarker values for protein density.

Outlier exclusions:

All dietary intake and biomarker variables were logarithmically transformed before analysis so that their distributions within each study/gender were approximately normal, as confirmed by visualization of quantile-quantile (QQ) plots. For each variable, extreme outlying values were excluded if the log value was less than Q1 – 2×(Q3-Q1) or greater than Q3 + 2×(Q3-Q1), where Q1 was the 25th percentile of the distribution of log values and Q3 was the 75th percentile. Overall less than 1% of all observations were so excluded, and the largest % of excluded variables for any variable in any study was 2.9% (DLW values in NPAAS).

Drift in repeated 24HR reports:

The 24HR energy data were assessed with regard to a possible drift over repeat assessments. The only study that showed marked drift was Energetics, which included 8 repeat assessments. It was decided to include only the second, third and fourth 24HR assessments in this study, because the first assessment with this non-interviewer-assisted instrument was thought to be on the participants' learning curve and because reported levels declined after the fourth assessment.

Analysis of 24HR reports:

Table 3 reports the geometric means of intakes based on the first administration of the instrument, except for Energetics where the second administration of the 24HR was used since it was thought that the participants in this study experienced a learning curve with its completion.

Web Appendix 2

Linear Mixed Model used for Meta-Analysis of Attenuation Factors

The attenuation factors were estimated through the linear mixed model for the biomarker:

$$M_{kij} = T_{ki} + e_{kij} = \lambda_{k0} + \lambda_{k1}Q_{ki} + u_{ki} + e_{kij} \ , \qquad (1)$$

where M_{kij} is the j^{th} observation of the biomarker for the i^{th} individual in study k, T_{ki} is the (unobserved) true usual intake of that individual, Q_{ki} is the self-report of that individual, u_{ki} is the random (unobserved) intercept for that individual, and e_{kij} is random within-person variation. The parameter λ_{k0} is the study-specific intercept and λ_{k1} is the study-specific attenuation factor. It is assumed that the random terms are independent normally distributed with mean 0, and study-specific variances.

Calibration (prediction) equations were estimated using a linear mixed model with additional predictors Z, as follows:

$$M_{kij} = T_{ki} + e_{kij} = \lambda_{k0} + \lambda_{k1}Q_{ki} + \lambda_2 Z_{ki} + u_{ki} + e_{kij}, \qquad (2)$$

where Z_{ki} is a vector of variables representing relevant personal characteristics of the i^{th} individual in the k^{th} study, and λ_2 is a vector of coefficients that is common to all the studies.

Meta-analyses of reporting bias were also performed on the basis of linear models, with M_{kij} in model (1) replaced by $Q_{ki} - \overline{M}_{ki.}$ (where $\overline{M}_{ki.}$ is the individual's mean log biomarker performed in the main study), and the terms $\lambda_{k1}Q_{ki} + u_{ki}$ on the right hand side omitted. For analyzing personal characteristics associated with reporting bias, the term $\lambda_2 Z_{ki}$ was added to the right had side.

Web Appendix 3

Statistical method of estimating correlation of reported intake with truth and R-squared for calibration equations, adjusting for within-person biomarker variation

The general model for calibration (prediction) of true intake is:

$$T_{ki} = \lambda_{k0} + \lambda_{k1}Q_{ki} + \lambda_2 Z_{ki} + u_{ki} \, , \qquad (1)$$

where T_{ki} is the (unobserved) true intake value of the i^th individual in study k, Q_{ki} is that individual's value of a self-report instrument, Z_{ki} is a vector of personal characteristics for that individual and u_{ki} is an unobserved random intercept for that individual.

The parameters of model (1) are estimated through the linear mixed model for the biomarker:

$$M_{kij} = T_{ki} + e_{kij} = \lambda_{k0} + \lambda_{k1}Q_{ki} + \lambda_2 Z_{ki} + u_{ki} + e_{kij} \, ,$$

(2)

where M_{ij} is the jth observation of the biomarker for individual i, and e_{ij} is random within-person variation (see Appendix 1).

For estimating the correlation of reported intake with truth, we use model (2) setting λ_2=0. The correlation with true intake for the k^th study is then given by

$$\operatorname{corr}(T, Q) = \sqrt{\frac{\lambda_{k1}^2 \sigma_{Qk}^2}{\lambda_{k1}^2 \sigma_{Qk}^2 + \sigma_{uk}^2}} \, ,$$

and σ_{Qk}^2 is estimated as the sample variance of Q in study k, while λ_{k1}, and σ_{uk}^2 are estimated from the model.

For estimating the multiple R^2 value for the general model (1), R^2 can be written as

$$R^2 = \operatorname{corr}^2\{T_k, E(T_k \mid Q_k, Z_k)\} = \frac{\lambda_k^T \Sigma_k \lambda_k}{\lambda_k^T \Sigma_k \lambda_k + \sigma_{uk}^2} \, , \qquad (3)$$

where $\lambda_k = (\lambda_{k1}\ \lambda_2)^T$, and Σ_k is the covariance matrix of (Q_k, Z_k). Parameters λ_{k1}, λ_2 and σ_{uk}^2 are estimated using the full model (2), while Σ_k is estimated as the sample covariance of (Q, Z) in study k.