Use of short-term dietary data for the estimation of usual intake

Victor Kipnis

Biometry, National Cancer Institute, USA

Acknowledgments

I am part of the Measurement Error Working Group at the U.S. National Cancer Institute.

We have been working together for 8+ years to address important issues in nutritional surveillance and nutritional epidemiology

Surveillance Measurement Error Group

Susan Krebs-Smith, Dennis Buckman, Raymond Carroll, Kevin Dodd, Laurence Freedman, Patricia Guenther, Victor Kipnis, Douglas Midthune , Amy Subar, Janet Tooze

NATIONAL CANCER INSTITUTE

TEXAS A&M UNIVERSITY

USDA

WAKE FOREST UNIVERSITY SCHOOL of MEDICINE THE BOWMAN GRAY CAMPUS

Dietary Assessment

- There are many methods to measure dietary intake both long-term (FFQ) and short-term (24HR, Food records, Biomarkers)

- More accurate short-term measurements have been often used in national surveys (e.g., 24HRs in US NHANES)

Dietary Assessment (2)

- Use of short-term measurements for estimating “usual” (long-term average) intakes is associated with numerous challenges

– even if otherwise precise, measurements contain substantial **within-person error** due to short-term variation in intake

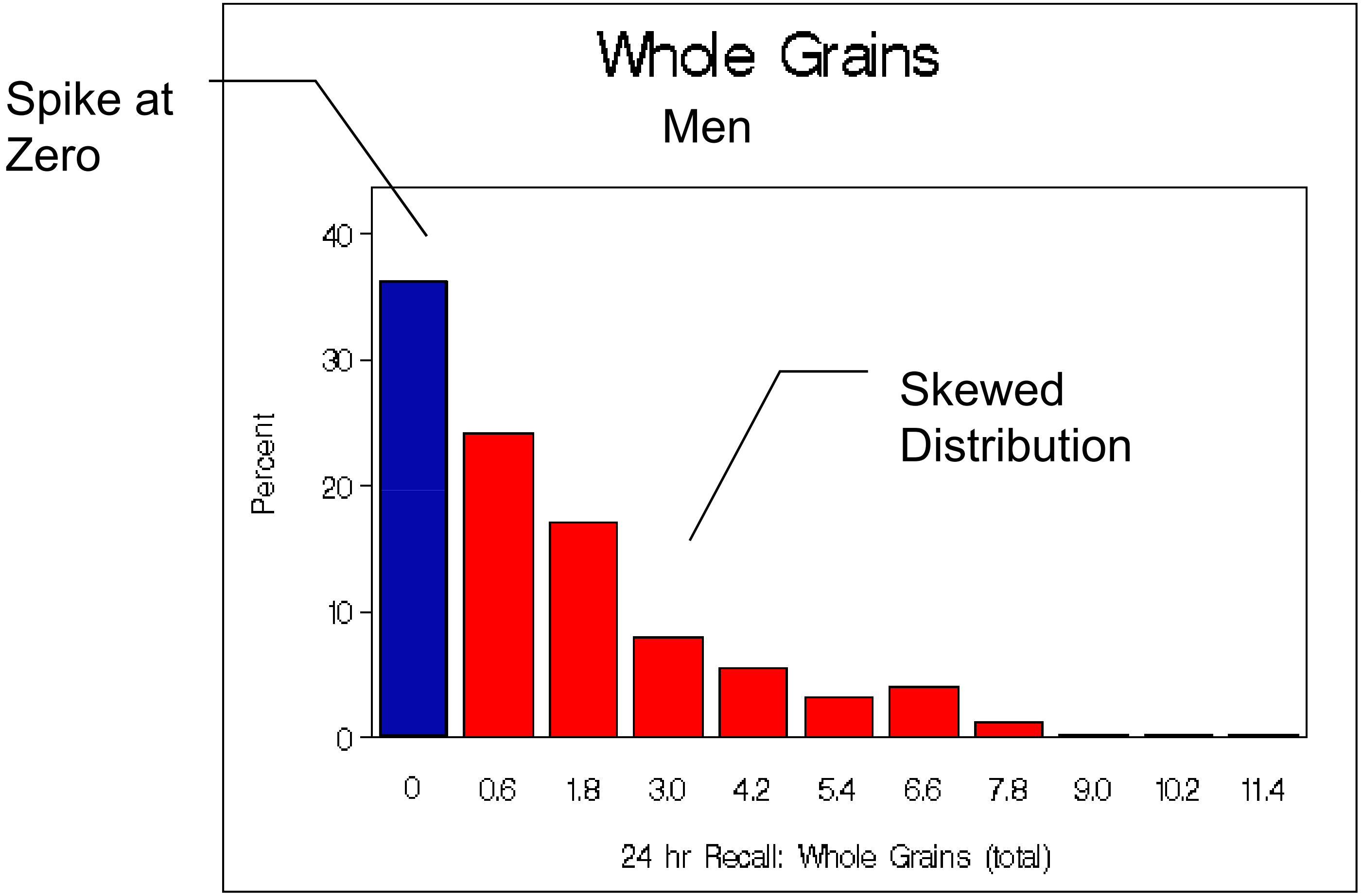
– measurements often have **skewed to the right** distributions

– measurements of *episodically consumed components* (i.e., not consumed daily by most) are **zero-inflated**

– dietary components are usually **mutually correlated** requiring **multivariate modeling**

Episodically consumed components

- Problem: short-term measurements have spike at zero and skewed to the right distribution of positive intake



Motivating example: HEI-2005 Scores in NHANES

- NHANES is US National Survey involving 2 24HRs
- HEI-2005 is scoring system based on a priori chosen dietary recommendations for densities of 6 episodically & 6 daily consumed dietary components:
  - total fruit; whole fruit; total veggies; dark green and orange veggies & legumes (DOL); total grains; whole grains; milk; meat & beans; oil; saturated fat; sodium; solid fats, alcoholic beverages and added sugars (SoFAAS)
- Higher scores indicate greater compliance with dietary guidelines (healthier diet)
- Total score (sum of individual scores) is on a scale 0 to 100 and estimating its distribution requires multivariate modeling

Modeling Assumption

- In what follows, we will consider studies where dietary assessment is done with repeat short-term measurements of *p* dietary components, of which first *m* are episodically and last $p-m$ are daily consumed
- **Main assumption**: for person *i*, repeat *j*, short-term measurement $R_{k,ij}$ of the *k*-th component is unbiased for its true usual intake
- The developed methodology is demonstrated using 24HR

New Multivariate Model (1)

- Methodology is a multivariate extension of the NCI method (Tooze et al., 2006; Kipnis et al. 2009; Tooze et al., 2010)
- Generically, ***X*** denotes model covariates (e.g., socio-economic characteristics, demographics), and ***β*** denotes population-based covariate effects (*fixed effects)*
- Generically, ***u*** denotes person-specific *random effects* representing part of within-person mean not explained by covariates
- Finally, ***ε*** denotes within-person variation in repeat measurements

New Multivariate Model (2)

- For an episodically consumed component $k=1,\ldots,m$, specify the two-part model as follows
  - Part I: Consider a *mixed effects* latent variable

$$\tilde{R}_{kij} = \boldsymbol{\beta}'_{2k-1,X}\mathbf{X}_i + u_{2k-1,i} + \varepsilon_{2k-1,ij}$$

The *fact of consumption* during period *j* is specified as $\tilde{R}_{kij} > 0$

  - Part II: Given consumption, consider Box-Cox transformation

$$g(v;\lambda) = (v^{\lambda} - 1) / \lambda$$

such that the *transformed amount* follows the *mixed effects linear model*

$$g_{R_k}\left(R_{kij}\right) = \boldsymbol{\beta}'_{2k,X}\mathbf{X}_i + u_{2k,i} + \varepsilon_{2k,ij}$$

New Multivariate Model (3)

- For a daily consumed component $k = m+1, \dots, p$, consider a Box-Cox transformation such that the *transformed amount* follows the *mixed effects linear model*

$$g_{R_k}\left(R_{kij}\right) = \boldsymbol{\beta}^{t}_{m+k,X}\mathbf{X}_{i} + u_{m+k,i} + \varepsilon_{m+k,ij}$$

New Multivariate Model (4)

- Person-specific random effects and within-person errors are specified as

$$\mathbf{u}_i = \left(u_{1i}, \ldots, u_{2m+p,i}\right)^t \sim N\left(\mathbf{0}; \boldsymbol{\Sigma}_{\mathbf{u}}\right)$$

$$\boldsymbol{\varepsilon}_{ij} = \left(\varepsilon_{1ij}, \ldots, \varepsilon_{2m+p,ij}\right)^t \sim N\left(\mathbf{0}; \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\right)$$

- For identifiability

$$\mathrm{var}\left(\varepsilon_{2k-1,ij}\right) = 1,\ k = 1, \ldots, m$$

- Based on two-part model specification,

$$\mathrm{var}\left(\varepsilon_{2k-1,ij}, \varepsilon_{2k,ij}\right) = 0,\ k = 1, \ldots, m$$

New Multivariate Model (5)

- Allowing correlations among person-specific random effects induces correlations among *usual intakes* of daily consumed and episodic components
- Allowing correlations among within-person errors induces

  – correlations among *short-term positive intakes* of daily and episodically consumed components

  – correlations of *indicators of short-term consumption* (yes, no) for episodically consumed components among themselves and with *consumed positive amounts of other components*

New Multivariate Model: fitting

- Denoting model parameters by Θ, we have a highly nonlinear mixed effects model

$$\mathbf{R}_{ij} \equiv \left(R_{1,ij}, \ldots, R_{p,ij}\right)^{t} = \Re(\mathbf{X}_i, \mathbf{u}_i, \boldsymbol{\varepsilon}_{ij}; \boldsymbol{\theta})$$

  with many correlated latent variables and the patterned covariance matrix of within-person errors with structured zeros and ones
- Currently available software cannot handle such models
- The model is therefore fitted using Markov Chain Monte-Carlo technique
- We treat the method as if it were non-Bayesian, and get standard errors using Balanced Repeated Replication (BRR)

Estimating the Distribution of Usual Intakes

- Multivariate true usual intake is defined as the mean of $\mathbf{R}_{ij}$

$$T_i = \int \mathfrak{R}\left(\mathbf{X}_i, u_i, \varepsilon_{ij}; \boldsymbol{\theta}\right) f\left(\varepsilon_{ij} \mid \mathbf{X}_i, u_i; \boldsymbol{\theta}\right) d\varepsilon \equiv \mathfrak{T}(\mathbf{X}_i, \mathbf{u}_i; \boldsymbol{\theta})$$

  Individual realizations of $\mathbf{u}_i$ (and thus true intakes) remain unknown

- BUT generating $\tilde{\mathbf{u}}_{ib} \sim \mathrm{N}\left(0, \hat{\boldsymbol{\Sigma}}_{\mathbf{u}}\right),\ b = 1, \ldots, B,$ one can estimate the distribution of true usual intakes or its function, such as the total HEI score, by weighted (survey weights) empirical distribution of $\mathfrak{T}(\mathbf{X}_i, \tilde{\mathbf{u}}_i; \hat{\boldsymbol{\theta}})$

- Details in Zhang et al., 2011

Example: HEI-2005 Scores in NHANES

- Data: 2 24HRs from 2001-2004 NHANES for children aged 2-8
- The vector of covariates $\mathbf{X}_i$ included age, gender, a dummy variable indicating weekend (Friday, Saturday, or Sunday) or week day, and a dummy variable indicating 1st or 2nd recall
- Important questions:

– correlations among HEI-2005 component scores

– distribution of the HEI-2005 total score

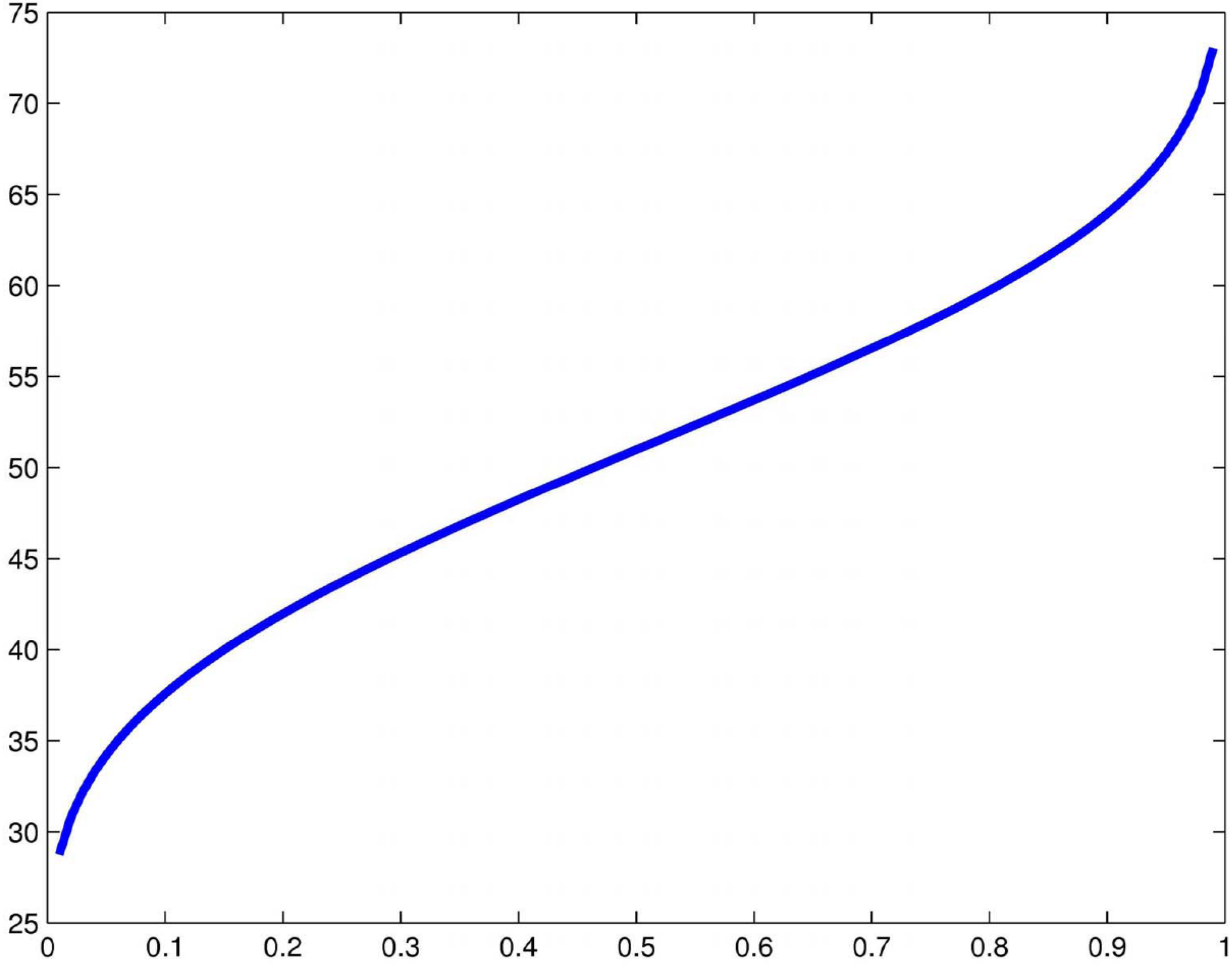
– distribution of HEI-2005 component scores among those with total score greater or smaller than, say, 50

– % of Americans exceeding percentiles for all 12 HEI-2005 component scores

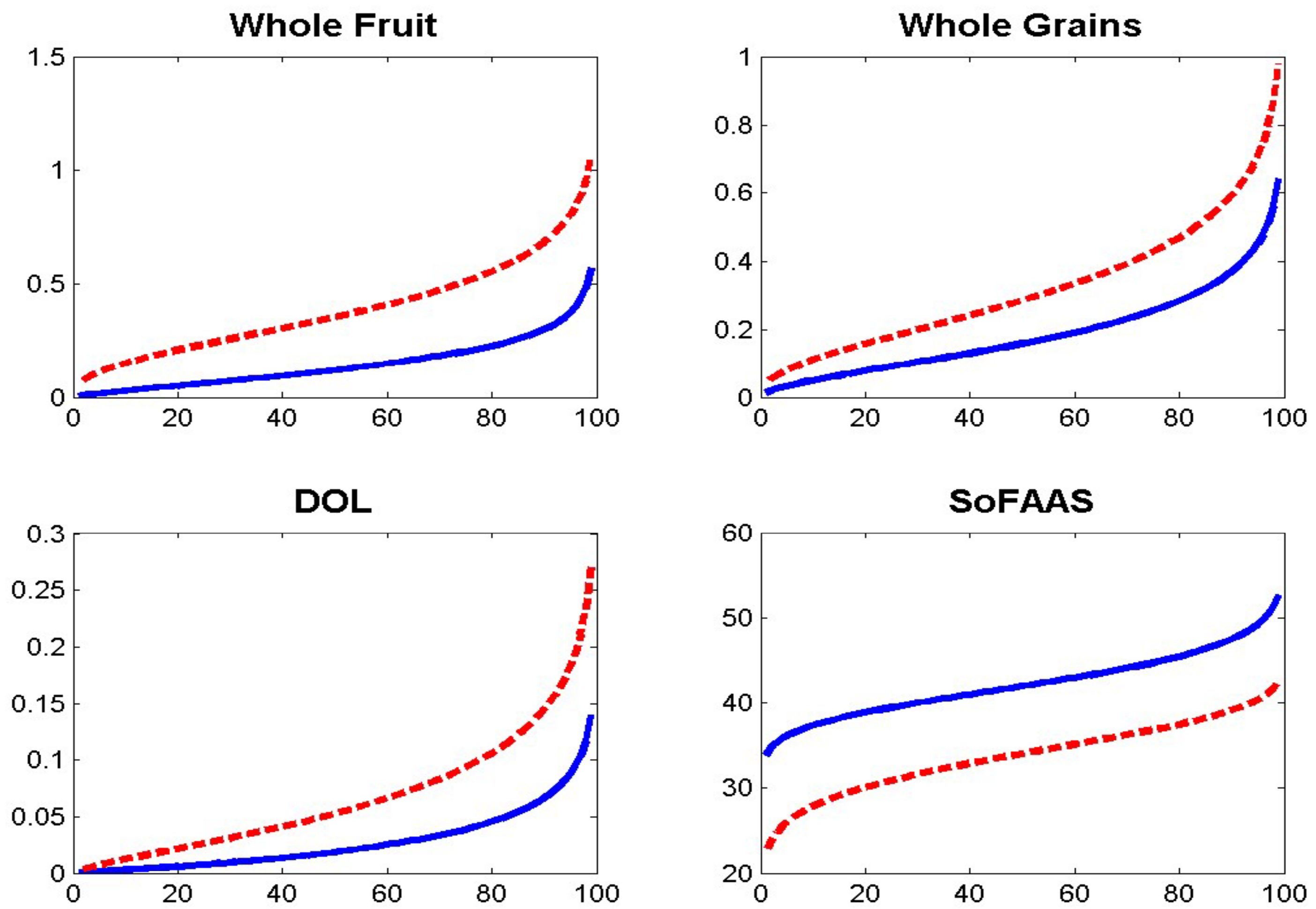
Estimated Correlations among HEI-2005 Components

	TF	WF	TG	WG	TV	DOL	Milk	Meat	Oil	SFat	Sod	SoFFAS
T Fruit	1	.75	-.09	.27	.05	.45	.15	.07	-.35	-.36	-.27	-.64
W Fruit		1	.04	.32	.13	.52	.09	.06	-.17	-.30	-.21	-.53
T Grns			1	.36	-.24	-.08	-.29	-.13	.44	-.36	.18	-.24
W Grns				1	-.22	.13	.16	-.17	-.11	-.31	-.14	-.46
T Veg's					1	.48	-.11	.53	-.08	.07	.44	-.16
DOL						1	.15	.25	-.09	-.26	-.02	-.50
Milk							1	-.37	-.22	.20	-.28	-.21
Meat								1	-.03	-.09	.41	-.21
Oil									1	-.06	.11	.04
Sat Fat										1	.09	.45
Sodium											1	.04
SoFFAS												1

Estimated percentiles of the HEI-2005 Total Score

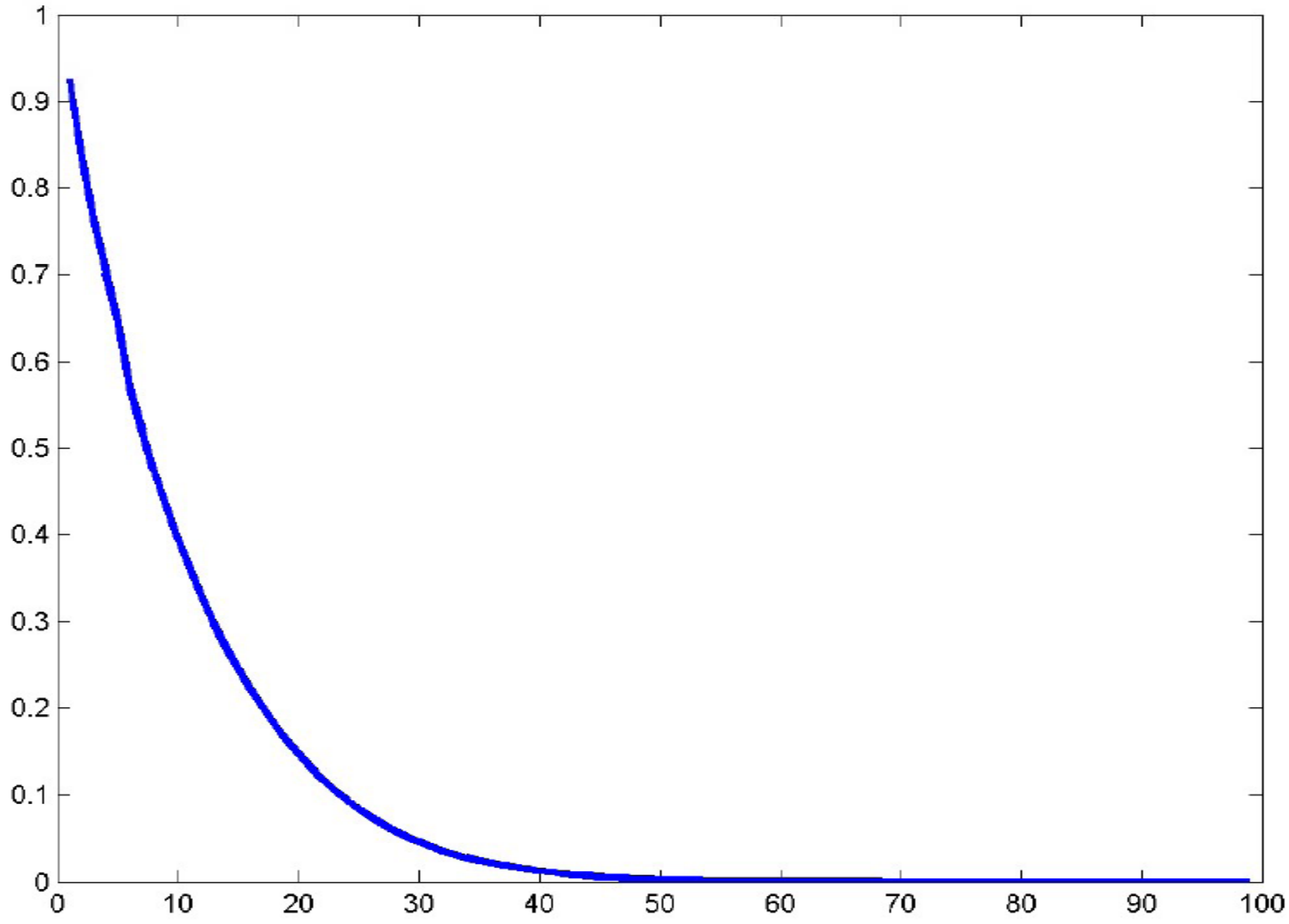


Estimated percentiles of HEI-2005 Components



Total Score ≤ 50 (Solid Blue);
Total Score > 50 (Dashed Red)

Estimated probabilities of exceeding k-th percentile on all 12 HEI-2005 components



Modeling Lifetime Exposure

- Let *c* denote survey cycle during which usual intake is assumed to be constant
- Using repeat short-term measurements during cycle *c*, true usual intake (average units per day) is modeled as

$$T_{ic} = \int \Re\left(\mathbf{X}_{ic}, u_{ic}, \varepsilon_{icj}; \boldsymbol{\theta}_c\right) f\left(\varepsilon_{icj} \mid \mathbf{X}_{ic}, u_{ic}; \boldsymbol{\theta}_c\right) d\varepsilon \equiv \mathfrak{T}_c(\mathbf{X}_{ic}, \mathbf{u}_{ic}; \boldsymbol{\theta}_c)$$

- Lifetime intake (up until cycle *C*) is defined as

$$T_i = \sum_{c=1}^{C} \mathfrak{T}_c(\mathbf{X}_{ic}, \mathbf{u}_{ic}; \boldsymbol{\theta}_c) \times t_c$$

where t_c denotes number of days in cycle *c*

Modeling Life-time Exposure (2)

- Ideal situation: everyone in the survey is followed up from cycle 1 to cycle *C*
- Modeling correlations among components of person-specific random effects $\{\mathbf{u}_{ic}\}$, one can estimate the distribution of life-time cumulative intake until cycle C by generating $\{\tilde{\mathbf{u}}_{ic}\}$ from the estimated multivariate normal distribution and using the resulting empirical weighted distribution of

$$T_i = \sum_{c=1}^{C} \mathfrak{T}_c(\mathbf{X}_{ic}, \tilde{\mathbf{u}}_{ic}; \hat{\boldsymbol{\theta}}_c) \times t_c$$

Modeling Life-time Exposure (3)

- Realistic situation: consecutive surveys sample different people in different cycles
- Data for modeling correlations among $\{\mathbf{u}_{ic}\}$ do not exist
- One way out: assuming that, although $\mathrm{var}(\mathbf{u}_{ic})$ may change from cycle to cycle, person-specific random effects for each person represent the same quantile in the distribution
- Modeling the variance of person-specific random effect in cycle *c* as

$$\sigma^2_{\mathbf{u}_{ic}} = \sigma^2_{u_c} \exp\{\gamma_c^t \mathbf{X}_{ic}\}$$

person-specific random effects could be generated as

$$\tilde{\mathbf{u}}_{ic} = \sigma_{u_{ic}} z_i,\ z_i \sim N(0,1)$$

Discussion

- New methodology addresses main challenges for simultaneous modeling of usual intakes of episodically and daily consumed dietary components using short-term unbiased assessment:

− in any short-term period, binary *indicators of episodic consumption* are allowed to be *correlated* among themselves and with *consumed positive amounts of other components*

− in any short-term period, all *daily consumed* and *positive amounts of episodically consumed components* are allowed to be mutually *correlated*

− all *usual dietary intakes* are allowed to be mutually *correlated*

Discussion (2)

- Marginal models for both episodically and daily consumed dietary components are the same as in the NCI univariate approach
- The model contains covariates and therefore could be used for subgroup analysis
- Frequentist treatment of MCMC methodology allows one to use BRR-based estimates of uncertainty
- In principle, the model could be extended to describe lifetime usual intake