National Cancer Institute

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
National Institutes of Health

measurement ERROR webinar series

Assessing diet-health relationships with FFQ: focus on episodically-consumed dietary components

Victor Kipnis, PhD
National Cancer Institute, USA

---

measurement ERROR webinar series

This series is dedicated to the memory of

Dr. Arthur Schatzkin

In recognition of his internationally renowned contributions to the field of nutrition epidemiology and his commitment to understanding measurement error associated with dietary assessment.

Presenters and Collaborators

Sharon Kirkpatrick
Series Organizer

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

Wake Forest School of Medicine · The Gertner Institute · STATISTICS TEXAS A&M UNIVERSITY

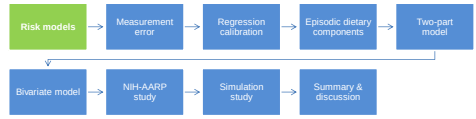
Learning objectives

- Review statistical risk models for evaluating diet-health relationships in nutritional epidemiology
- Learn application of regression calibration to correct for FFQ measurement error using repeat short-term reference measurements in a substudy
- With focus on episodically-consumed dietary components, learn application of a new methodology to carry out regression calibration in risk models with energy-adjusted dietary covariates

---

Outline

- Risk models in nutritional epidemiology
- Dietary measurement error
- Regression calibration
- Modeling episodically-consumed dietary components
  - Two-part model and its extensions
  - Three-part model for episodic component & energy
- Example: NIH-AARP Diet & Health Study
- Simulation study
- Summary & discussion



RISK MODELS IN NUTRITIONAL EPIDEMIOLOGY

Risk models in nutritional epidemiology

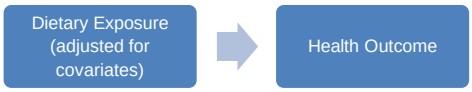
Types of epidemiologic studies

- Animal experiments
- Ecological studies
- Cross-sectional studies
- Case-control studies
- **Cohort studies** (main focus here)
- Randomized prevention trials

Risk models in nutritional epidemiology

Risk models: exposure

- We consider studies that relate:



- Dietary exposure thought to be most relevant is usual (long-term average) daily dietary intake
- Health outcome examples: continuous (e.g., blood pressure), binary (event, no event), time to event (survival analysis)

Risk models in nutritional epidemiology

Risk models: general description

- Notations:
  - Y - health outcome
  - $\mathbf{T} = (T_1, \ldots, T_p)^t$ - vector of dietary components
  - $\mathbf{Z} = (Z_1, \ldots, Z_q)^t$ - vector of adjusting covariates
  - $r(Y \mid \mathbf{T},\mathbf{Z})$ - outcome risk function
  - $\eta(\mathbf{T},\mathbf{Z}; \boldsymbol{\alpha})$ - covariate-based predictor ($\boldsymbol{\alpha}$ is a vector of parameters)
- Risk model: $r(Y \mid \mathbf{T},\mathbf{Z}) = \eta(\mathbf{T},\mathbf{Z}; \boldsymbol{\alpha})$

Risk models in nutritional epidemiology

Risk models: examples

- Common risk models:
  - **Linear regression** for continuous outcome (e.g., blood pressure, cholesterol level)
  - **Logistic regression** for binary outcome (event, no event)
  - **Cox regression** for survival analysis (time to event)

Risk models in nutritional epidemiology

Risk models: risk function (1)

- **Linear regression**
  - Outcome: Y - continuous variable (e.g., blood pressure, cholesterol level, etc.)
  - Risk function: conditional expected value (mean) given covariates, i.e.,

$$r(Y \mid \mathbf{T}, \mathbf{Z}) = \mathrm{E}\left(Y \mid \mathbf{T}, \mathbf{Z}\right)$$

Risk models in nutritional epidemiology

Risk models: risk function (2)

- **Logistic regression**
  - Outcome: binary variable $Y = \begin{cases} 1 & \text{if event} \\ 0 & \text{if no event} \end{cases}$
  - Risk function: logit of the probability of event (log odds of event) conditional on covariates, i.e.,

$$r(Y \mid \mathbf{T}, \mathbf{Z}) = \log \frac{\mathrm{P}(Y = 1 \mid \mathbf{T}, \mathbf{Z})}{1 - \mathrm{P}(Y = 1 \mid \mathbf{T}, \mathbf{Z})}$$

Risk models: risk function (3)

- **Cox regression**
  - Outcome: $Y = t$ (time to event)
  - Risk function: log of the hazard function $h(t \mid \mathbf{T}, \mathbf{Z})$ conditional on covariates, i.e.,

$$r(Y \mid \mathbf{T},\mathbf{Z}) = \log h(t \mid \mathbf{T},\mathbf{Z})$$

Risk models: risk predictor (1)

- Commonly used predictor is a linear function of covariates

$$\eta(\mathbf{T},\mathbf{Z};\boldsymbol{\alpha}) = \alpha_o + \sum_{k=1}^{p} \alpha_{T_k} T_k + \sum_{l=1}^{q} \alpha_{Z_l} Z_l$$

  - Note: α_0 is a constant in linear and logistic regressions and $\alpha_0 = h_0(t)$ (baseline hazard) in Cox regression
- Convenient but doesn't always provide a good fit
  - Example: orange vegetables vs. lung cancer in NIH-AARP Diet and Health Study (to be discussed later)

Risk models: risk predictor (2)

- A more flexible risk model specifies predictor as linear over transformed covariates

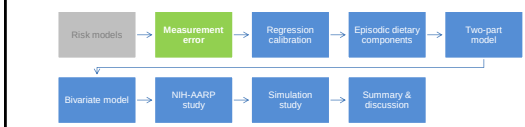
$$\eta(\mathbf{T},\mathbf{Z};\boldsymbol{\alpha}) = \alpha_o + \sum_{k=1}^{p} \alpha_{T_k} T_k^* + \sum_{l=1}^{q} \alpha_{Z_l} Z_l^*$$

  where for any variable V, $V^* = g_v(V)$ denotes its transformed value
- Popular transformations include power functions $g_v(V) = V^\gamma$ and logarithm $g_v(V) = \log V$

Risk models: risk predictor (3)

- Risk model: $r(Y \mid \mathbf{T},\mathbf{Z}) = \alpha_o + \sum_{k=1}^{p} \alpha_{T_k} T_k^* + \sum_{l=1}^{q} \alpha_{Z_l} Z_l^*$
- Slope α_{T_k} represents the effect of exposure T_k
  - Due to exposure transformation, this effect depends not only on change in exposure (case of linear predictor on original scale) but also on its initial value
  - Effect of changing exposure from T_{k0} to $T_{k1} = T_{k0} + \Delta T_k$ on risk $r(Y \mid \mathbf{T},\mathbf{Z})$ is

$$\alpha_{T_k}\left[g_{T_k}(T_{k0} + \Delta T_k) - g_{T_k}(T_{k0})\right]$$



MEASUREMENT ERROR

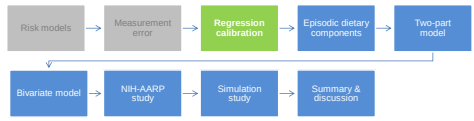
Dietary measurement error (1)

- Problem in nutritional epidemiology: true usual intakes are **unknown** and measured with error
- Assessment of diet in nutritional epidemiology is commonly done by food frequency questionnaire (FFQ) querying diet over a specified time period (usually 1 year)
- FFQ is known to contain substantial measurement error, both **random** and **systematic**

Measurement error

Dietary measurement error (2)

- Generally, fitting risk models to error-prone measured dietary exposures **Q** leads to:
  - Bias (often attenuation) of estimated exposure effect
  - Reduced power to detect exposure effect
  - In theory, invalid significance test for the main exposure (multiple error-prone covariates)
- Most popular method for correcting for dietary measurement error: **regression calibration**



REGRESSION CALIBRATION

Regression calibration

Regression calibration (1)

- Main assumption: measurement error is **nondifferential** with respect to health outcome, i.e., provides no additional information about the outcome beyond that in true diet
- This assumption may be justified in cohort studies where diet is usually assessed before outcome is known, but not necessarily in case-control studies due to possible recall bias when cases report their past diet differently from non-cases

Regression calibration

Regression calibration (2)

- **Regression calibration (RC)**: each mismeasured covariate in a risk model is replaced with its best predictor

$$T_k^{*P}(\mathbf{X}) = \mathrm{E}(T_k^* \mid \mathbf{X}),\ k = 1,\ldots,p,$$

given vector **X** that includes all observed error-prone covariates **Q** and error-free covariates **Z**

- RC leads to (approximately) true regression slopes, i.e., true covariate effects

$$r(Y \mid \mathbf{X}) = \tilde{\alpha}_0 + \sum_{k=1}^{p} \alpha_{T_k} T_k^{*P}(\mathbf{X}) + \sum_{l=1}^{q} \alpha_{Z_l} Z_l^*$$

Regression calibration

Regression calibration (3)

- In absence of true intakes, each predictor T_i^{*P} is estimated in a substudy (called calibration study) using (often repeat) reference measurements R_{ij}
- Requirement: reference measurements may contain error but should be unbiased for true individual usual intake, i.e., for person i, repeat j

$$\mathbb{E}(R_{ij} \mid i) = T_{ki}$$

- Regression calibration predictor is given by

$$T_i^{*P} = \mathbb{E}\left[g_T(T_i) \mid \mathbf{X}_i\right] = \mathbb{E}\left[g_T\left\{\mathbb{E}(R_{ij} \mid i)\right\} \mid \mathbf{X}_i\right]$$

and its estimation requires a model for R_{ij}

Regression calibration

Regression calibration (4)

- Ideal reference measurements of dietary intakes:
  - Short-term recovery biomarkers (unfortunately, only few are known: DLW for energy, UN for protein, UK for potassium)
- Reference measures in practice:
  - More detailed short-term self-reports such as multiple-day food records or repeat 24-hour dietary recalls (24HRs)

Regression calibration (5)

- Methodology below is developed for any correct reference measurement
- This methodology is demonstrated using 24HR (reference instrument in many important dietary cohorts)
  - Short-term reference period is 1 day
- **Working assumption**: 24HR is unbiased in reporting individual's true usual dietary intake
  - Implications of possible biases in 24HR are discussed at the end

Regression calibration (6)

- Given a measurement error model for reference measurements, regression calibration predictor is a function of covariates $\mathbf{X}_i$ and model parameters $\boldsymbol{\theta}$

$$T_i^{*P} = \mathbb{E}\left[g_T \left\{ \mathbb{E}\left(R_{ij} \mid i \right) \right\} \mid \mathbf{X}_i \right] = \mathfrak{T}(\mathbf{X}_i; \boldsymbol{\theta})$$

- Parameters are estimated in a calibration substudy as $\hat{\boldsymbol{\theta}}$
- Regression calibration predictor is then estimated for all subjects in the main study as

$$\hat{T}_i^{*P} = \mathfrak{T}(\mathbf{X}_i; \hat{\boldsymbol{\theta}}),\ i = 1, \ldots, N$$

Regression calibration (7)

Regularly-consumed dietary components

- **Ideal world**: the classical measurement error model

$$R_{ij} = T_i + \varepsilon_{ij},\ \varepsilon_{ij} \sim N\left(0; \sigma_\varepsilon^2\right)$$

where the regression of T_i on $\mathbf{X}_i$ is linear, i.e.,

$$T_i = \beta_0 + \boldsymbol{\beta}_X^t \mathbf{X}_i + u_i,\ u_i \sim N\left(0; \sigma_u^2\right)$$

- The measurement error model for reference measurements is thus specified as

$$R_{ij} = \beta_0 + \boldsymbol{\beta}_X^t \mathbf{X}_i + u_i + \varepsilon_{ij}$$

Regression calibration (8)

- Measurement error model

$$R_{ij} = \beta_0 + \boldsymbol{\beta}_X^t \mathbf{X}_i + u_i + \varepsilon_{ij}$$

is a **mixed effects linear model** which includes

- **fixed** (in this case linear) **effect** of covariates defined by the population-level parameters (β_0, $\boldsymbol{\beta}_X$)
- **random effect** u_i representing part of within-person mean not explained by covariates; it is person-specific but randomly varies across people
- **within-person random error** ε_{ij} representing short-term variation

Regression calibration (9)

- Regression calibration predictor on a transformed scale is given by

$$T_i^{*P} = \mathbb{E}\left[g_T \left\{ \mathbb{E}\left(R_{ij} \mid i \right) \right\} \mid \mathbf{X}_i \right] = \mathbb{E}\left[g_T \left\{ \beta_0 + \boldsymbol{\beta}_X^t \mathbf{X}_i + u_i \right\} \mid \mathbf{X}_i \right]$$

- For the risk model with predictor on original scale, conditional expectation above exists in closed forms, so that the regression calibration predictor is a linear function of covariates

$$T_i^P = \beta_0 + \boldsymbol{\beta}_X^t \mathbf{X}_i$$

Regression calibration (10)

- **Real world:** often within-person random error in R_{ij} depends on individual mean and has a skewed distribution, violating classical model assumptions
- Usual remedy: transformation to a scale where classical model is a good approximation, i.e.,

$$g_R\left(R_{ij}\right) = \beta_0 + \boldsymbol{\beta}_X^t \mathbf{X}_i + u_i + \varepsilon_{ij},\ u_i \sim N\left(0; \sigma_u^2\right),\ \varepsilon_{ij} \sim N\left(0; \sigma_\varepsilon^2\right)$$

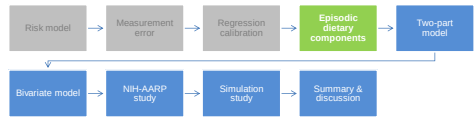
- Regression calibration predictor is then given by

$$\begin{aligned} T_i^{*P} &= \mathbb{E}\left[g_T \left\{ \mathbb{E}\left(R_{ij} \mid i \right) \right\} \mid \mathbf{X}_i \right] \\ &= \mathbb{E}\left[g_T \left\{ \mathbb{E}\left(g_R^{-1} \left\{ \beta_0 + \boldsymbol{\beta}_X^t \mathbf{X}_i + u_i + \varepsilon_{ij} \right\} \mid i \right) \right\} \mid \mathbf{X}_i \right] \end{aligned}$$

Regression calibration

Linear regression calibration

- Linear regression calibration approximation (LRC): commonly used in nutritional epidemiology
- **Working assumptions**:
  - There is a scale where R_{ij}^* is well approximated by the linear mixed effects model and the risk model's predictor, is linear i.e., $g_T(.) = g_R(.) = g(.)$
  - On this scale, approximately $\mathbb{E}\left(R_{ij}^* \mid i\right) \approx T_i^*$
- Then $T_i^{*P} = \mathbb{E}\left[\mathbb{E}\left(R_{ij}^* \mid i\right) \mid \mathbf{X}_i\right] = \beta_0 + \boldsymbol{\beta}_X' \mathbf{X}_i$
- We will see later that LRC may fail to provide a good approximation for nonlinear models



EPISODIC DIETARY COMPONENTS

Episodic dietary components

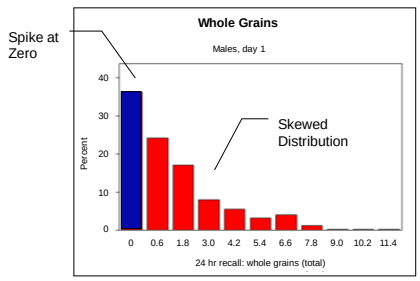
Episodic dietary components (1)

- Our focus: **episodically-consumed dietary components** i.e., those that are not consumed daily by nearly everyone (but are eventually consumed in the long run)
- Examples:
  - Many foods (fish, red meat, whole grains, dark green or orange vegetables, etc.)
  - Some nutrients (vitamin A or B12)

Episodic dietary components

Episodic dietary components (2)

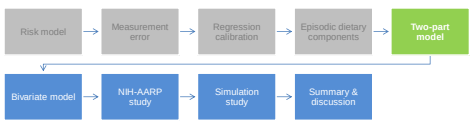
- Example: typical short-term report



Episodic dietary components

Episodic dietary components (3)

- Short-term reference measurements for episodically-consumed dietary components
  - Are **semicontinuous variables** with **excess zeros** and often **skewed** to the right positive values
  - Even if otherwise precise, contain substantial within-person **measurement error** due to short-term variation in intake



TWO-PART MODEL

Two-part model

Two-part model (1)

- Goal: specifying a measurement error model for semicontinuous reference measurements
- Main idea: modeling a semicontinuous variable as the result of **two** distinct, although generally correlated **processes**:
  - One determines whether the variable takes positive or zero value
  - Other specifies its positive value

Two-part model

Two-part model (2)

- Two-part model – first proposed by Cragg (1971) and intensively studied in econometrics and (later) in biostatistics
  - Part I – logit/probit regression specifying the probability of positive values
  - Part II – linear regression specifying log-transformed positive values

Two-part model

Two-part model (3)

- Extended to longitudinal data by Olsen & Schafer (2001) and Tooze et al. (2002) by introducing **mixed effects** two-part model with:
  - **Fixed effects** that are defined by a function of covariates with population-level parameters
  - **Random effects** that represent part of within-subject mean not explained by covariates; it is subject-specific but randomly varies across subjects
  - Within-subject **random errors** in positive values representing longitudinal variation

Two-part model

Two-part model (4)

- Longitudinal two-part mixed effects model:
  - Part I – mixed effects logistic regression specifying the probability of positive values
  - Part II – mixed effects linear regression for log-transformed positive values
  - Both parts are linked by allowing correlated person-specific random effects and overlapping covariates
- For model identifiability, need at least 2 repeat observations on at least a subsample of subjects

Two-part model

Two-part model (5)

- New methodology (NCI method) further extended the longitudinal two-part mixed effects model for short-term reference measurements of episodically-consumed dietary components by:
  - Including Box-Cox family of transformations of positive values (to allow flexibility in addressing skewedness)

$$g(v;\gamma_v)=\begin{cases}(v^{\gamma_v}-1)/\gamma_v & \text{if } \gamma_v \neq 0\\ \log(v) & \text{if } \gamma_v = 0\end{cases}$$

  - Allowing for within-subject random measurement error

Two-part model

Two-part model (6)

- Two-part NCI model:

$$\mathbb{P}(R_{ij}>0\,|\,i)=H\left(\beta_{10}+\boldsymbol{\beta}_1^t\mathbf{X}_{1i}+u_{1i}\right),\ H(v)=\left(1+e^{-v}\right)^{-1}$$
$$g(R_{ij};\gamma_R\,|\,R_{ij}>0)=\beta_{20}+\boldsymbol{\beta}_2^t\mathbf{X}_{2i}+u_{2i}+\varepsilon_{2ij},\ \varepsilon_{2ij}\sim N\left(0,\sigma_{\varepsilon_2}^2\right)$$

where:

$$\mathbf{u}_i=(u_{1i},u_{2i})^t\sim N\left(\mathbf{0};\boldsymbol{\Sigma}_{\mathbf{u}}\right),\ \boldsymbol{\Sigma}_{\mathbf{u}}=\begin{pmatrix}\sigma_{u1}^2 & \sigma_{u1,u2}\\ \sigma_{u1,u2} & \sigma_{u2}^2\end{pmatrix}$$

- Part I specifies the probability of consumption & part II specifies the consumption amount; both parts are linked by allowing correlated person-specific random effects and overlapping covariates

Two-part model

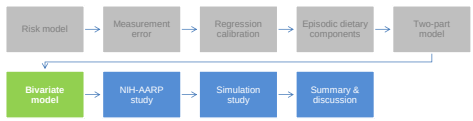
Two-part model (7)

- For a **single** episodically consumed dietary component, using two repeat 24HRs in US NHANES Survey as the main dietary-assessment instrument, NCI method was applied to:
  - Estimating the distribution of usual intake and its characteristics (Tooze et al, JADA, 2006)
  - Estimating relationships of usual intake with health outcome (Kipnis et al, Biometrics, 2009)

Two-part model

Two-part model (8)

- Goal: extending NCI methodology for adjusting diet-health relationships for FFQ measurement error when the risk model includes **several** dietary components
  - In many cases, regression calibration can be applied to error-prone covariates in a risk model one by one
  - But there is a problem with dietary risk models due to energy adjustment



BIVARIATE MODEL

Bivariate model

Bivariate model (1)

- To understand effects of dietary composition, epidemiologists usually consider risk models with energy-adjusted dietary covariates such as:
  - **Density**, i.e., ratio of usual intake of interest to usual energy intake (focus here)
  - **Residual** from regressing usual intake of interest on usual energy intake
- Energy-adjusted risk models also include energy
- Since many dietary intakes are correlated with energy intake, estimation of RC predictor requires modeling episodic component and energy together

Bivariate model

Bivariate model (2)

- **Observed data** in calibration sub-study: for person i, time period j
  - R_{Fij}, R_{Eij} - short-term reference measurements of episodic dietary component F and energy E
  - $\mathbf{X}_i$ - vector of observed covariates, including FFQ-reported intakes $\mathbf{Q}_i$ and error-free covariates $\mathbf{Z}_i$
  - Indicator variable of reference consumption in period j for episodic component

$$I_{Fij} = I(R_{Fij} > 0) = \begin{cases} 1 & \text{if } R_{Fij} > 0 \\ 0 & \text{if } R_{Fij} = 0 \end{cases}$$

Bivariate model

Bivariate model (3)

- **Latent (unobserved) variables**:
  - T_{Fij}, T_{Eij} - true intakes of interest in period j
  - T_{Fi} - true usual intake of component of interest
  - T_{Ei} - true usual energy intake
  - $T_{Di} = T_{Fi} / T_{Ei}$ - true density intake of interest
- Additional latent variables: person-specific random effects and within-person random errors

Bivariate model (4)

- Energy intake is naturally specified as part II of the NCI model since energy is always consumed
- Allowing correlations between person-specific random effects in energy and episodic component models induces correlation between usual energy and episodic component intakes
- Allowing correlation between within-person errors in energy and part II of episodic component models induces correlation between energy and episodic component during short-term consumption period

Bivariate model (5)

- Additional model requirement:
  - During any short-term period, energy intake (continuous variable) should be allowed to be correlated with the indicator of consumption of dietary component of interest (binary variable)
- Original part I of the NCI model specifies a model for the probability to consume an episodic component but not for the indicator of short-term consumption
- To satisfy the above requirement, need to modify part I of the two-part NCI model

Bivariate model (6)

- **Modified part I**: to allow I_{Fij} & R_{Eij} to be correlated
  - Consider continuous **latent variable** in period *j*

$$\tilde{R}_{Fij} = \beta_{F10} + \boldsymbol{\beta}_{F1}^{t}\mathbf{X}_i + u_{F1i} + \varepsilon_{F1ij},$$

where $u_{F1i} \sim N\left(0, \sigma^2_{u_{F1}}\right)$, $\varepsilon_{F1ij} \sim N(0,1)$

  - Let $\tilde{R}_{Fij}$ underlie binary indicator of episodic component's reference consumption

$$I_{Fij} = 1 \Leftrightarrow \tilde{R}_{Fij} > 0$$

  - Allow ε_{F1ij} and within-person error in the model for reference energy intake R_{Eij} to be correlated

Bivariate model (7)

- The use of additional latent variable $\tilde{R}_{Fij}$ leads to the probability of consumption specified as the mixed effects **probit** model

$$\mathbb{P}(R_{Fij} > 0 \mid i) = \Phi\left(\beta_{F10} + \boldsymbol{\beta}_{F1}^{t}\mathbf{X}_i + u_{F1i}\right)$$

where Φ denotes the distribution function of the standard normal random variable

- Note: in part I of the original NCI model the probability of consumption is specified as the logit model without underlying continuous latent variable in period *j*

Bivariate model (8)

- **Part II**: the same as in the original NCI model, i.e., transformed consumption amount during period *j* is specified as linear mixed effects model:

$$g(R_{Fij}; \gamma_{R_F} \mid R_{Fij} > 0) = \beta_{20} + \boldsymbol{\beta}_{F2}^{t}\mathbf{X}_i + u_{F2i} + \varepsilon_{F2ij}$$

where:

$$u_{F2i} \sim N\left(0, \sigma^2_{u_{F2}}\right),\ \varepsilon_{F2ij} \sim N\left(0, \sigma^2_{\varepsilon_{F2}}\right)$$

- Part I and II are **linked** by using the same covariates and allowing person-specific random effects to be correlated

Bivariate model (9)

- **Model for Energy**: transformed energy amount for period *j* is specified as linear mixed effects model:

$$g(R_{Eij}; \gamma_{R_E}) = \beta_{E0} + \boldsymbol{\beta}_{E}^{t}\mathbf{X}_i + u_{Ei} + \varepsilon_{Eij}$$

where:

$$u_{Ei} \sim N\left(0, \sigma^2_{u_E}\right),\ \varepsilon_{Eij} \sim N\left(0, \sigma^2_{\varepsilon_E}\right)$$

- Model for energy is **linked** to both parts I and II for episodic component by using the same covariates and allowing:
  - person-specific random effects to be correlated
  - within-person errors ε_{Eij} to be correlated with ε_{F1ij} and ε_{F2ij}

Bivariate model

Bivariate model (10)

- Bivariate model is formally specified as:

$$R_{Fij} = I\left[\beta_{F10} + \boldsymbol{\beta}_{F1}^{t}\mathbf{X}_i + u_{F1i} + \varepsilon_{F1ij} > 0\right] \times g^{-1}\left(\beta_{F20} + \boldsymbol{\beta}_{F2}^{t}\mathbf{X}_i + u_{F2i} + \varepsilon_{F2ij}; \gamma_{R_F}\right)$$

$$R_{Eij} = g^{-1}\left(\beta_{E0} + \boldsymbol{\beta}_{E}^{t}\mathbf{X}_i + u_{Ei} + \varepsilon_{Eij}; \gamma_{R_E}\right)$$

where: $\mathbf{u}_i = (u_{F1i}, u_{F2i}, u_{Ei})^t \sim MVN\left(\mathbf{0}; \boldsymbol{\Sigma}_{\mathbf{u}}\right)$

$$\boldsymbol{\varepsilon}_{ij} = (\varepsilon_{F1ij}, \varepsilon_{F2ij}, \varepsilon_{Eij})^t \sim MVN\left\{\mathbf{0}; \begin{pmatrix} 1 & 0 & \sigma_{\varepsilon_{F1E}} \\ 0 & \sigma^2_{\varepsilon_{F2}} & \sigma_{\varepsilon_{F2E}} \\ \sigma_{\varepsilon_{F1E}} & \sigma_{\varepsilon_{F2E}} & \sigma^2_{\varepsilon_E} \end{pmatrix}\right\}$$

Bivariate model

Bivariate model (11)

- Denoting by $\boldsymbol{\theta}_F, \boldsymbol{\theta}_E$ model parameters for episodic component and energy, respectively, we have:

$$R_{Fij} = \Re_F(\mathbf{X}_i, \mathbf{u}_{Fi}, \boldsymbol{\varepsilon}_{Fij}; \boldsymbol{\theta}_F),\ R_{Eij} = \Re_E(\mathbf{X}_i, u_{Ei}, \varepsilon_{Eij}; \boldsymbol{\theta}_E)$$

- True usual intakes of episodic component and energy are expectations of those functions, i.e.,

$$T_{Fi} = \mathbb{E}\left\{\Re_F\left(\mathbf{X}_i, \mathbf{u}_{Fi}, \boldsymbol{\varepsilon}_{Fij}; \boldsymbol{\theta}_F\right) | i\right\} = \Im_F(\mathbf{X}_i, \mathbf{u}_{Fi}; \boldsymbol{\theta}_F),$$

$$T_{Ei} = \mathbb{E}\left\{\Re_E\left(\mathbf{X}_i, u_{Ei}, \varepsilon_{Eij}; \boldsymbol{\theta}_E\right) | i\right\} = \Im_E(\mathbf{X}_i, u_{Ei}; \boldsymbol{\theta}_E)$$

Bivariate model

Bivariate model (12)

- True episodic component density is given by:

$$T_{Di} \equiv \frac{T_{Fi}}{T_{Ei}} = \Im_D(\mathbf{X}_i, \mathbf{u}_i; \boldsymbol{\theta}_F, \boldsymbol{\theta}_E)$$

- Regression calibration predictor for T^*_{Di} is given by:

$$T^{*P}_{Di} \equiv \mathbb{E}\left[g\left\{T_{Di}; \gamma_{T_D}\right\} | \mathbf{X}_i\right] = \mathbb{E}\left[g\left\{\Im_D(\mathbf{X}_i, \mathbf{u}_i; \boldsymbol{\theta}_F, \boldsymbol{\theta}_E, \gamma_{T_D})\right\} | \mathbf{X}_i\right]$$

- Regression calibration predictor for T^*_{Ei} is given by:

$$T^{*P}_{Ei} \equiv \mathbb{E}\left[g\left\{T_{Ei}; \gamma_{T_E}\right\} | \mathbf{X}_i\right] = \mathbb{E}\left[g\left\{\Im_E(\mathbf{X}, u_E; \boldsymbol{\theta}_E, \gamma_{T_E})\right\} | \mathbf{X}_i\right]$$

Bivariate model

Bivariate model (13)

- Model parameters $\boldsymbol{\theta}_F, \boldsymbol{\theta}_E$ in the bivariate model are estimated by fitting the model in the calibration sub-study by MLE using NLMIXED SAS procedure
- For any given set of covariate transformations, by using estimated model parameters $\hat{\boldsymbol{\theta}}_F, \hat{\boldsymbol{\theta}}_E$, regression calibration predictors for transformed density and energy are calculated for each person in the main study
- The final set of covariate transformations is chosen to maximize the overall likelihood when fitting the risk model

NIH-AARP STUDY

NIH-AARP study

NIH-AARP diet and health study (1)

- Prospective cohort of 567,169 men & women aged 50-71 in 1995-96 with FFQ administered at baseline
- Calibration substudy of ~1000 men and 1000 women with 2 non-consecutive 24HRs
- Analysis: relationships in men between
  - red meat density & lung cancer
  - orange vegetables density & lung cancer

  adjusting for age, smoking, and energy intake
- Risk model: Cox regression on original and Box-Cox transformed scales with standard errors estimated by bootstrap to account for estimated RC predictors

NIH-AARP study

NIH-AARP diet and health study (2)

- Due to zeros, for linear regression calibration (LRC), used Box-Cox transformation with a shift parameter:

$$g(v;\gamma_v)=\begin{cases}\dfrac{(v+\delta)^{\gamma_v}}{\gamma_v} & \text{if } \gamma_v \neq 0\\ \log(v+\delta) & \text{if } \gamma_v = 0\end{cases}$$

- To assess covariate transformations, risk model fit was tested using cumulative martingales technique implemented in SAS (p <0.05 indicates poor fit)
- Compared FFQ-based analysis (no correction for measurement error) with corrections using linear regression calibration and regression calibration based on the bivariate model

NIH-AARP study

NIH-AARP diet and health study (3)

24HR reported consumption of orange vegetables and red meat

	Red Meat (g/day)	Orange Vegetables (cups/day)
Mean reported intake (s.e.)	82.8 (2.3)	0.14 (0.01)
Mean amount on consumption days (s.e.)	117.7 (2.6)	0.32 (0.01)
Mean probability to consume (s.e.)	0.70 (0.01)	0.44 (0.01)
% of subjects who consumed food:		
0 out of 2 days	14.7	33.5
1 out of 2 days	29.9	44.9
2 out of 2 days	55.4	21.6

NIH-AARP study

NIH-AARP diet and health study (4)

NIH-AARP Diet and Health study: Red meat intake and lung cancer risk in men; hazard ratios for red meat density between 10 & 60 g/1000 kcal

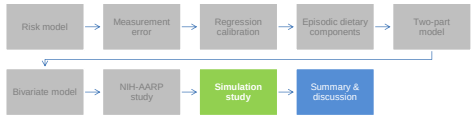
Risk Model: Measurement Error Correction Method	γ	δ	Estimated Log Hazard Ratio (s.e.)	Estimated Hazard Ratio (95% CI)	Risk model fit test p-value
Untransformed Intake:					
No correction for ME	1		0.225(0.040)	1.252(1.158,1.354)	0.041
RC (Bivariate model)	**1**		**0.409(0.075)**	**1.505(1.300,1.744)**	**0.130**
LRC	1		0.441(0.097)	1.554(1.285,1.880)	0.041
Transformed Intake:					
No correction for ME	0.4		0.248(0.046)	1.281(1.171,1.402)	0.082
RC (Bivariate model)	**1**		**0.409(0.077)**	**1.505(1.294,1.751)**	**0.130**
LRC (chosen scale)	0.4	0.001	0.321(0.155)	1.379(1.017,1.868)	0.076
LRC (fixed log scale)	0	0.001	0.113(0.096)	1.120(0.928,1.351)	0.564

NIH-AARP study

NIH-AARP diet and health study (5)

NIH-AARP Diet and Health study: Orange vegetable intake and lung cancer risk in men; hazard ratios for orange vegetable density from 0.02 to 0.10 cups/1000 kcal

Risk Model: Measurement Error Correction Method	γ	δ	Estimated Log Hazard Ratio (s.e.)	Estimated Hazard Ratio (95% CI)	Risk model fit test p-value
Untransformed Intake:					
No correction for ME	1		-0.076(0.021)	0.927(0.889,0.966)	<0.0001
RC (Bivariate model)	**1**		**-0.265(0.078)**	**0.767(0.658,0.894)**	**0.002**
LRC	1		-0.223(0.086)	0.800(0.676,0.947)	<0.0001
Transformed Intake					
No correction for ME	-0.3		-0.182(0.030)	0.834(0.786,0.884)	0.256
RC (Bivariate model)	**0.1**		**-0.380(0.089)**	**0.684(0.574,0.814)**	**0.060**
LRC (chosen scale)	-20	1	-0.593(0.146)	0.553(0.415,0.736)	0.202
LRC (log scale)	0	0.005	-0.387(0.107)	0.679(0.551,0.838)	0.022



SIMULATION STUDY

Simulation study

Simulation study (1)

- Main study: for 100 000 subjects generated FFQ and 1000 24HRs with distributions similar to those of orange vegetables and energy in NIH-AARP study
- Calibration substudy: for 1000 subjects used first 2 24HRs as reference measures
- True usual intakes: calculated as averages of 1000 24HRs; density intakes were calculated as ratios of true usual component to usual energy intakes
- Binary outcome: generated using logistic regression with Box-Cox transformed exposure

Simulation study

Simulation study (2)

- Goal: estimating log RR for increasing main exposure between 0.02 & 0.10 cups/1000 kcal
- Risk model: logistic regression on original and Box-Cox transformed scales with standard errors estimated by bootstrap
- Compared FFQ-based analysis (no correction for measurement error) with corrections using linear regression calibration and regression calibration based on the bivariate model

Simulation study

Simulation study: results (1)

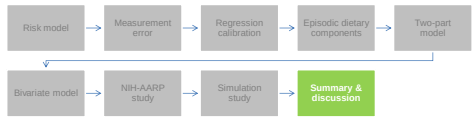
Results of the simulation study: mean, standard deviation and root mean squared error of estimated log odds ratio in logistic regression of disease on orange vegetable intake and energy

Sim	Measurement Error Correction Method	Mean γ (Mean δ)	Mean Log OR (s.e.)	Standard Deviation	RMSE
1	True parameters	1	-0.4		
	No correction for ME	0.42	-0.242 (0.002)	0.025	0.160
	RC (Bivariate model)	**0.85**	**-0.417(0.005)**	**0.078**	**0.080**
	LRC (original scale)	1	-0.436(0.008)	0.118	0.123
	LRC (log scale)	0 (0.08)	-0.557(0.008)	0.108	0.191
	LRC (chosen scale)	-0.61 (0.18)	-0.543(0.007)	0.104	0.177
2	True parameters	0.1	-0.4		
	No correction for ME	0.02	-0.207(0.002)	0.024	0.194
	RC (Bivariate model)	**0.10**	**-0.416(0.007)**	**0.093**	**0.094**
	LRC (original scale)	1	-0.285(0.006)	0.088	0.145
	LRC (log scale)	0 (0.006)	-0.342(0.006)	0.083	0.101
	LRC (chosen scale)	-2.18 (0.14)	-0.375(0.009)	0.121	0.124

Simulation study

Simulation study: results (2)

- In theory, for risk models on **original** scale, LRC is approximately consistent, BUT leads to **finite sample biases** due to unaccounted excess zeros
- For risk models on a **transformed** scale, LRC may not perform well because
  - Trying to find a scale where both calibration and risk models have linear predictors and reference measurements are unbiased often leads to poor approximations
  - Applying LRC on the original scale by definition leads to a misspecified risk model



SUMMARY & DISCUSSION

Summary

Summary (1)

- Developed methodology addresses major challenges for bivariate modeling of short-term reference intakes of an episodic component & energy by allowing during any short-term period:
  - Energy intake to be correlated with the indicator of episodic component consumption
  - Energy intake to be correlated with consumption amount

Summary

Summary (2)

- Developed methodology allows for rigorous regression calibration to correct for nondifferential covariate measurement error in rather flexible risk models with multiple dietary exposures that:
  - Include energy-adjusted dietary components
  - Include covariates on transformed scales
- Simulations indicate that the developed method performs substantially better than conventional linear regression calibration

Discussion

Discussion (1)

- Focus here: episodically-consumed dietary components that are eventually consumed in the long run
- What about never consumers?
  - Model could be extended to include never consumers
  - Depending on dietary component and a reference instrument, it may require more than 2 repeat reference measurements (e.g., 4-6 with 24HR-reported fish intake, Kipnis et al, Biometrics 2009)

Discussion

Discussion (2)

- Developed methodology is based on the important assumption that a short-term reference instrument is **unbiased** for true usual dietary intake on individual level
- In considered applications, such instrument was 24HR
- Studies with recovery biomarkers (DLW for energy, UN for protein, UK for potassium) demonstrate some bias in 24HR, suggesting possible biases in reporting other dietary components

Discussion

Discussion (3)

- Recent publication (Freedman et al, JNCI 2011) based on OPEN biomarker study suggests that, in spite of biases, using 24HR as a reference to correct for FFQ measurement error on average leads to better results than FFQ-based analysis with no correction for measurement error
- Using more precise short-term reference instruments, such as 24HR, for correcting for FFQ measurement error is a step forward toward better understanding of diet-health outcome relationships

Discussion

Discussion (4)

- Using reference measurements to calibrate FFQ (approximately) removes bias but does not fully restore the **power** to detect a relationship, which is lost due to measurement error
- Even bias correction may not be reliable if the attenuated effect is too small (weak signal problem)
- One can do better by using more precise short-term instruments (e.g., web-based ASA24) as the main dietary-assessment method and/or combine different instruments
- The corresponding methodologies will be presented in webinars 10-12

QUESTIONS & ANSWERS

Moderator: Kevin Dodd

Please submit questions using the *Chat* function

National Cancer Institute

measurement ERROR webinar series

Next Session

Tuesday, November 15, 2011
10:00-11:30 EST

Estimating usual intake distributions for multivariate dietary variables

Raymond J. Carroll
Texas A&M University

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
National Institutes of Health