Correcting for Measurement Error in Reporting of Episodically Consumed Foods When Estimating Diet-Disease Relationships

Victor Kipnis, National Cancer Institute, USA
Raymond Carroll, Texas A&M U, USA
Laurence Freedman, Gertner Institute, Israel
Douglas Midthune, National Cancer Institute, USA

OUTLINE

- Introduction: impact of dietary measurement error
- Regression calibration correction
- Challenges to analysis of episodically consumed foods
- Regression calibration model
- Simulation study:
  - whole grains vs colorectal cancer in men
  - fish vs colorectal cancer in men
- Example: red/processed meat vs lung cancer in NIH-AARP Diet & Health Study
- Discussion

Impact of Measurement Error

- Food Frequency Questionnaire (FFQ) is instrument of choice in most studies in nutritional epidemiology
- FFQ is known to contain substantial measurement error, random and systematic
- Typically measurement error causes two things:
  – bias in the estimated exposure effect (often leading to flattened or attenuated true slope in disease model)
  – loss of statistical power to detect exposure effect

Impact of Measurement Error

- Disease model: for disease outcome D, vector $\boldsymbol{T} = (T_1, ..., T_K)^t$ of true usual intakes, and vector $\boldsymbol{Z} = (Z_1, ..., Z_L)^t$ of covariates

$$\mathbb{E}(D|\boldsymbol{T}, \boldsymbol{Z}) = m(\alpha_0 + \boldsymbol{\alpha}_T^t \boldsymbol{T} + \boldsymbol{\alpha}_Z^t \boldsymbol{Z})$$

  where $m^{-1}(.)$ is link function (e.g., logit)

- Main assumption: errors in reported intakes $\boldsymbol{Q}$ are non-differential with respect to outcome D, i.e.

$$\mathcal{F}(D|\boldsymbol{T}, \boldsymbol{Q}, \boldsymbol{Z}) = \mathcal{F}(D|\boldsymbol{T}, \boldsymbol{Z})$$

  – Example: conditional distribution of reported intakes given true intakes is the same among cases and controls

Regression Calibration

- Disease model

$$\mathbb{E}(D|\boldsymbol{T},\boldsymbol{Z}) = m\big(\alpha_0 + \boldsymbol{\alpha}_T^t\boldsymbol{T} + \boldsymbol{\alpha}_Z^t\boldsymbol{Z}\big)$$

- Regression calibration: to a very good approximation

$$\mathbb{E}(D|\boldsymbol{Q},\boldsymbol{Z}) = m\big(\alpha_0 + \boldsymbol{\alpha}_T^t\mathbb{E}(\boldsymbol{T}|\boldsymbol{Q},\boldsymbol{Z}) + \boldsymbol{\alpha}_Z^t\boldsymbol{Z}\big)$$

- Intuition: substitution for unknown vector $\boldsymbol{T}$ its best prediction given the reported intakes $\boldsymbol{Q}$ and covariates $\boldsymbol{Z}$

Regression Calibration

- In absence of gold standard, regression calibration predictors $\mathbb{E}(T_k|\boldsymbol{Q},\boldsymbol{Z})$, $k = 1, ..., K$ are estimated using short-term reference measurements
- For continuous intake, reference measurements are required to satisfy classical error model

$$R_{ij} = T_i + \epsilon_{ij}$$

  where errors ϵ_{ij} are additive, independent of true intake, errors in FFQ, and each other
- Then regression calibration predictor can be estimated as

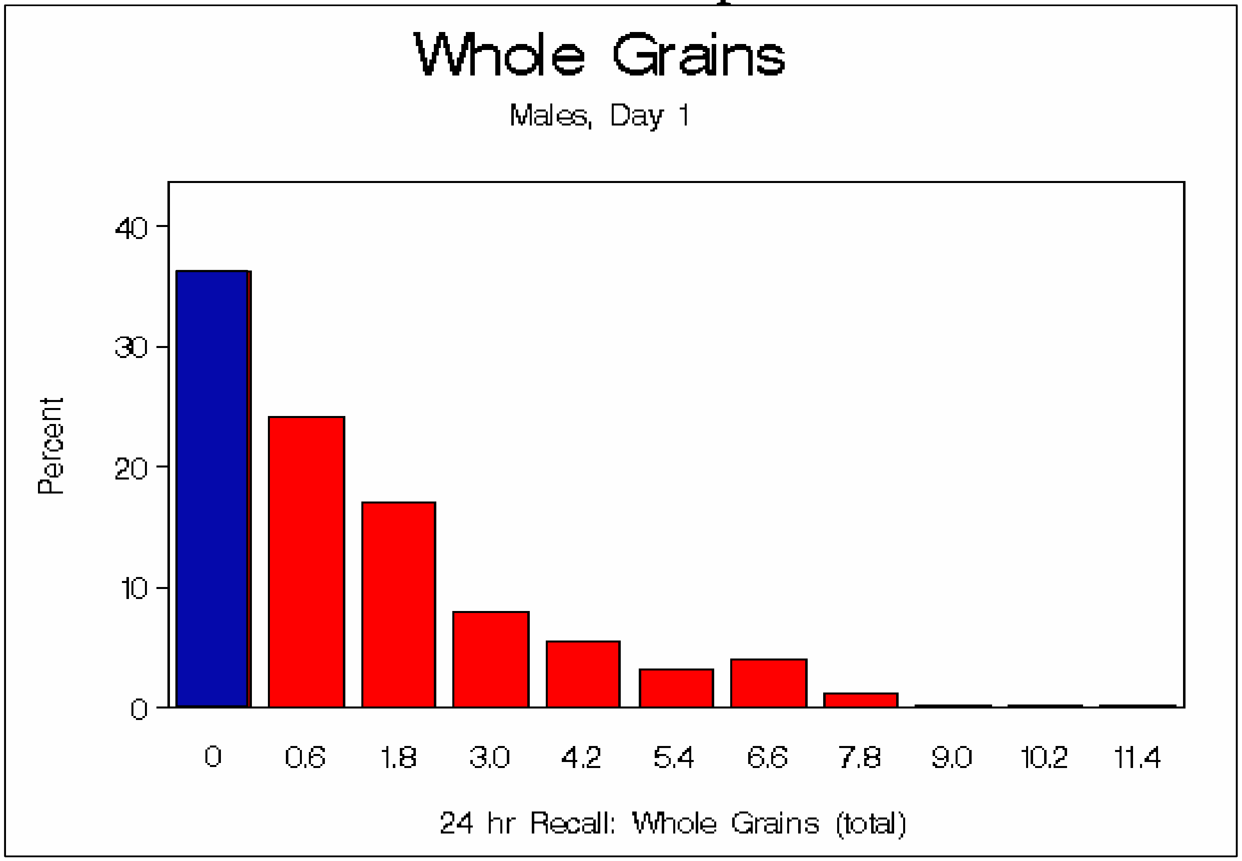
$$\mathbb{E}(R_{ij}|\boldsymbol{Q}_i,\boldsymbol{Z}_i) = \mathbb{E}(T_i|\boldsymbol{Q}_i,\boldsymbol{Z}_i)$$

Regression Calibration

- Ideal reference measure
  - short-term 'recovery' biomarker
- Reference measure in reality
  - more extensive short-term dietary-assessment method such as 24HR or diary
- 24HR is of special interest because it is used in 2 largest cohorts, AARP and EPIC
- Distributions of nutrient intakes are typically rather skewed: classical error model for reference measure may not hold
- Remedy: transformation to a scale where classical error model holds

Intake of Episodically Consumed Foods

- Problem: short-term reference measure (e.g., 24HR) has spike at zero and skewed distribution of positive intake



Statistical Model: true usual intake

- For person i, day j, and intake T_{ij} of interest, let

$$p_i = \mathbb{P}(T_{ij} > 0|i)$$

denote *probability* to consume on any given day

- Let

$$A_i = \mathbb{E}(T_{ij}|i; T_{ij} > 0)$$

denote usual consumption *amount*

- Then usual intake, defined as $T_i = \mathbb{E}(T_{ij}|i)$, is given by

$$T_i = \mathbb{E}(T_{ij}|i; T_{ij} > 0) \times \mathbb{P}(T_{ij} > 0|i) = p_i A_i$$

Statistical Model: assumptions for reference instrument

- Conditional on (transformed) $\boldsymbol{X}_i = (\boldsymbol{Q}_i^t, \boldsymbol{Z}_i^t)^t$

$$\mathbb{P}(R_{ij} > 0|\boldsymbol{X}_i) = \mathbb{P}(T_{ij} > 0|\boldsymbol{X}_i)$$

- For a monotone transformation g(.) reference amount on transformed scale has classical measurement error

$$g(R_{ij}|R_{ij} > 0) = \mu_{R_i} + \epsilon_{ij}\,,\ \epsilon_{ij}\ \sim N(o, \sigma_\epsilon^2)$$

- Reference amount is unbiased on transformed scale:

$$\mathbb{E}\{g(R_{ij})|i, R_{ij} > 0\} = g(A_i)$$

Statistical Model: part I

- **Part I – Probability to consume**
  - Logistic regression (mixed model)

$$\mathbb{P}(R_{ij} > 0|\boldsymbol{X}_i) = \mathbb{P}(T_{ij} > 0|\boldsymbol{X}_i)$$

$$= H(\beta_{01} + \boldsymbol{\beta}_{X1}^t \boldsymbol{X}_i + u_{1i})$$

where

$H(v) = (1 + e^{-v})^{-1}$ is logistic function

$u_{1i} \sim N(0, \sigma_{u_1}^2)$ is person-specific random effect allowing person's value to differ from that defined by covariates

Statistical Model: part II

- **Part II – Amount on consumption day**
  - Linear regression (mixed model) on transformed scale

$$g(R_{ij}|R_{ij} > 0; \boldsymbol{X}_i) = \mu_{R_i} + \epsilon_{ij}$$

$$= \beta_{02} + \boldsymbol{\beta}^t_{X2}\boldsymbol{X}_i + u_{2i} + \epsilon_{ij}$$

where

$g(v) = (v^\theta - 1)/\theta$ – Box-Cox transformation

$u_{2i} \sim N(o, \sigma^2_{u_2})$ – person-specific random effect

$\epsilon_{ij} \sim N(o, \sigma^2_\epsilon)$ – within-person random error

Statistical Model

- **Two-part model**

$$\mathbb{P}(R_{ij} > 0 | \boldsymbol{X}_i) = H(\beta_{01} + \boldsymbol{\beta}_{X1}^t \boldsymbol{X}_i + u_{1i})$$
$$g(R_{ij} | R_{ij} > 0; \boldsymbol{X}_i) = \beta_{02} + \boldsymbol{\beta}_{X2}^t \boldsymbol{X}_i + u_{2i} + \epsilon_{ij}$$

- **Link**

$$(u_{1i}, u_{2i})^t \sim N(\mathbf{0}, \boldsymbol{\Sigma}),\ \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{u_1}^2 & \rho_{u_1,u_2}\sigma_{u_1}\sigma_{u_2} \\ & \sigma_{u_2}^2 \end{pmatrix}$$

  – person-specific random effects are correlated

  – covariates can be (partially) shared

Regression Calibration Model

- True usual intake

$$T_i = H(\beta_{01} + \boldsymbol{\beta}^t_{X1}\boldsymbol{X}_i + u_{1i}) \times g^{-1}(\beta_{02} + \boldsymbol{\beta}^t_{X2}\boldsymbol{X}_i + u_{2i})$$

- Regression-calibration predictor for transformed $h(T_i)$

$$\mathbb{E}[h\{H(\beta_{01} + \boldsymbol{\beta}^t_{X1}\boldsymbol{X}_i + u_{1i})g^{-1}(\beta_{02} + \boldsymbol{\beta}^t_{X2}\boldsymbol{X}_i + u_{2i})\}|\boldsymbol{X}_i]$$

- Linear regression calibration:
  – Monte Carlo estimation of regression calibration predictors by generating $\hat{\boldsymbol{u}} = (\hat{u}_{1i}, \hat{u}_{2i})^t \sim N(\mathbf{0}, \hat{\boldsymbol{\Sigma}})$, using estimated parameters $(\hat{\beta}_{01}, \hat{\beta}_{02}, \hat{\boldsymbol{\beta}}_{X1}, \hat{\boldsymbol{\beta}}_{X2})$ to calculate $h(\widehat{T}_i)$ and regressing $h(\widehat{T}_i)$ on $\boldsymbol{X}_i$

EATS: Design

- Men and women 20-70 years
- Nationally representative sampling of 12,615 telephone numbers
- Approximately 1600 recruited
- Four 24HRs, one in each season
- After one year: DHQ about past year
- 886 respondents completed four 24HRs and DHQ

Simulation Study

- Idea: simulate data that are similar to reported intake of *whole grains* and *fish* in EATS
  – transform FFQ using best Box-Cox transformation to approximate normality
  – fit two-part model relating 4 24HRs to transformed FFQ, Q^*, and estimate model parameters
  – generate $\boldsymbol{u}_i = (u_{1i}, u_{2i}) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, $i = 1, ..., 20,000$
  – generate $Q_i^* \sim N(\mu_{Q^*}, \sigma^2_{Q^*})$, $i = 1, ..., 20,000$
  – generate two 24HRs for $1,000$ subjects in calibration study

$$R_{ij} = \begin{cases} 0 \text{ with pr} = 1 - p_i,\ p_i = H(\beta_{01} + \beta_{Q1} Q_i^* + u_{1i}) \\ g^{-1}(\beta_{02} + \beta_{Q2} Q_i^* + u_{2i} + \epsilon_{ij}) \text{ with pr} = p_i \end{cases}$$

Simulation Study

- Generate $T_i = H(\beta_{01} + \beta_{Q1} Q_i^* + u_{1i}) g^{-1}(\beta_{02} + \beta_{Q2} Q_i^* + u_{2i})$ $i = 1, ..., 20,000,$ and transform to $h(T_i)$ using best Box-Cox transformation to approximate normality
- Generate binary outcome variable for colorectal cancer in men

$$\mathbb{P}(D_i = 1 | T_i) = H\{\alpha_0 + \alpha_1 h(T_i)\}$$

where $\alpha_1$represents $log\, RR = 0.5$ for increasing exposure from $10\,\%$ to $90\ \%$ of the true exposure distribution and $\alpha_0 = \ -3.05$ which corresponds to probability of 3% for a 60 y old man to get disease in general population within 10 years

Simulation Study

- Comparison of 4 different methods:
  - true exposure on transformed scale
  - FFQ-reported exposure on transformed scale
  - "conventional" regression calibration approach by using mean of 2 24HRs on transformed scale as reference instrument
  - suggested regression calibration
- Since different methods lead to fitting risk model on different scales, RR is always calculated for the given increase in intake from a to $b = a + \Delta$, where a is equal to 10th percentile and b is equal to 90th percentile of true exposure on original scale

Simulation Study: Results

- True log RR for increase in *whole grain* intake from 0.25 to 2.85 pyramid servings/day is equal to -0.74

	Method			
log RR	True exposure	FFQ	Naive RC	New RC
Mean (s.e.)	-0.74 (.008)	-0.47 (.008)	-0.59 (.01)	-0.75 (.012)
St. dev.	0.110	0.107	0.134	0.174
RMSE	0.110	0.290	0.201	0.174

Simulation Study: Results

- True log RR for increase in *fish* intake from 0.064 to 1.39 oz/day is equal to -0.69

	Method			
log RR	True exposure	FFQ	Naive RC	New RC
Mean (s.e.)	-0.70 (.008)	-0.51 (.008)	-0.97 (.017)	-0.71 (.012)
St. dev.	0.105	0.106	0.234	0.166
RMSE	0.105	0.216	0.361	0.167

NIH-AARP Diet & Health Study

- Prospective cohort of 567,169 men & women aged 50-71 in 1995-96
- FFQ administered at baseline
- Calibration substudy of ~ 1000 men and ~1000 women with 2 24HRs and additional FFQ
- Analysis: association between red/processed meat and lung cancer for 349,148 men using Cox regression
- Confounders: age, BMI, smoking, physical activity, education, non-red/non-processed meat, fruit, total energy

NIH-AARP Diet & Health Study: Meat & Lung CA

	Method		
	FFQ	Naive RC	New RC
Red Meat			
HR(10% – 90%)	1.22	1.29	1.38
Bootstrap 95% CI	(1.10; 1.35)	(1.05; 1.61)	(1.06; 1.81)
Processed meat			
HR(10% – 90%)	1.18	1.22	1.34
Bootstrap 95% CI	(1.09; 1.28)	(1.11; 1.36)	(1.16; 1.56)

Discussion

- New method addresses all of the challenges for modeling usual intake of foods and overcomes the limitations of conventional regression calibration
  - Models intake as the product of probability to consume and consumption amount
  - Allows for skewed distribution of reference consumption amount by transforming to a scale with classical error model
  - Allows probability and amount to be correlated
  - Uses rigorous regression calibration approach

Discussion

- Method is based on important assumptions that reference instrument correctly specifies probability of short-term fact of consumption and that, on appropriate scale, it follows classical measurement error for consumption amount
- Studies with unbiased biomarker (DLW) for energy expenditure have found bias in reporting of energy intake on 24HR
  - suggests systematic misreporting of at least some foods
- For foods reported with bias on 24HR, correction for measurement error using 24HR as reference instrument will be biased as well