National Cancer Institute

measurement ERROR webinar series

Assessing diet-health relationships using a short-term unbiased dietary instrument: focus on risk models with multiple dietary components

Victor Kipnis, PhD
National Cancer Institute, USA

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

Slide 1

Hello and welcome to the 12th session in the Measurement Error Webinar Series. I'm Sharon Kirkpatrick, a nutrition researcher at the U.S. National Cancer Institute, and it's my pleasure to host today's webinar, our final session of the series, in which we'll continue with our focus on advanced and emerging topics.

Please note that the webinar is being recorded so that we can make it available on our Web site for future reference. All phone lines have been muted and will remain that way throughout the webinar. Please use the Chat feature if you'd like to submit a question for the question and answer period that will follow the presentation. And, lastly, you can find the slides for today's presentation on the Web site that has been set up for series participants. The URL is available in the Notes box at the top left of the screen.

Now I'd like to introduce Dr. Victor Kipnis, our presenter for today. Victor is a mathematical statistician in the Biometry Research Group, Division of Cancer Prevention, at the National Cancer Institute of the United States. Victor's research focus is on the structure of dietary measurement error, its effects on study results, and methods of adjusting for it in nutritional epidemiology and surveillance. In today's session, Victor will discuss assessing diet and health relationships using a short-term unbiased dietary instrument, with a focus on risk models with multiple dietary components. Victor.

Thank you very much, Sharon, and welcome everyone. So here is my title. *(V. Kipnis)*

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.

Slide 2

And this webinar, like all other webinars in this series, is dedicated to the memory of Dr. Arthur Schatzkin, a friend, a former colleague, who was very much interested in this topic of dietary measurement error.

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

Slide 3

And here are the names of people who have been involved in this project.

Learning objectives

- Review statistical risk models for evaluating diet-health relationships in nutritional epidemiology
- Learn application of regression calibration to correct for measurement error in a single dietary exposure when diet is assessed by repeat administration of a short-term unbiased instrument
- Learn application of a new methodology to carry out regression calibration in risk models with multiple dietary components (some of which are episodically consumed) measured by repeat administration of a short-term unbiased instrument

Slide 4

The learning objectives: firstly, I will briefly review statistical risk models for evaluating diet-health relationships that are used in nutritional epidemiology. Then, we will learn application of regression calibration to correct for measurement error in a single dietary exposure in the case when diet is assessed by repeat administration of a short-term unbiased instrument. Then, we will learn the new methodology of carrying out the same regression calibration but in risk models with multiple dietary components, some of which are episodically consumed and some of which are consumed daily. Again, this is a case when we have repeat administration of a short-term unbiased instrument.

Outline

- Risk models in nutritional epidemiology
- Dietary measurement error
- Regression calibration using repeat short-term unbiased measurements:
  - Single dietary component
  - New methodology for multivariate extension
- Simulation study
- Summary & discussion

Slide 5

Outline: We will briefly talk about risk models, dietary measurement error, and then in more detail about regression calibration using repeat short-term unbiased instruments. Then, we will discuss the results of the simulation study and the summary and discussion. And I would like to say up front that some of the material that we're going to discuss you've heard before, some concepts and some models, and even in the area of regression calibration in general, although its application to the case when the measurements are short-term unbiased measurements of dietary intakes is a new one, especially in the multivariate case.

RISK MODELS IN NUTRITIONAL EPIDEMIOLOGY

Slide 6

All right, so risk models….

Types of epidemiologic studies

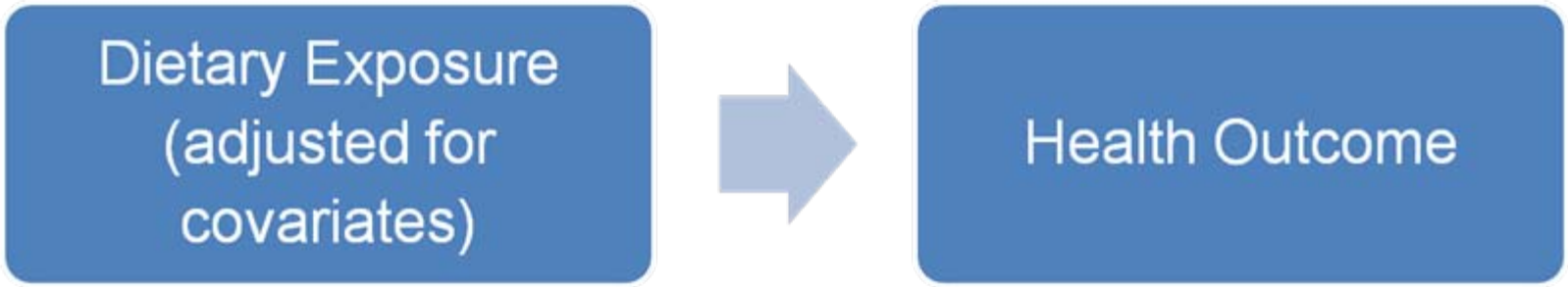
- Animal experiments
- Ecological studies
- Cross-sectional studies
- Case-control studies
- **Cohort studies** (main focus here)
- Randomized prevention trials

Slide 7

Here are the types of epidemiologic studies, and I will focus on cohort studies. This will be my main focus and in a few moments I will explain why.

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)

Slide 8

And what we're going to do is we're going to relate dietary exposure, adjusted for the covariates, to a health outcome. And that exposure thought to be most relevant is what we call usual, or long-term, average daily dietary intake. And it could be univariate, but in most cases, of course, it's multivariate.

Health outcome examples are continuous outcomes—for example, blood pressure; binary outcomes—for example, disease or no disease, or event or no event; and time to event in survival analysis.

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
  - $\eta(\mathbf{T},\mathbf{Z}; \alpha)$ - covariate-based predictor (α is a vector of parameters)
- Risk model: $r(Y \mid \mathbf{T},\mathbf{Z}) = \eta(\mathbf{T},\mathbf{Z}; \alpha)$

Slide 9

Here are my notations. Y will denote the health outcome. T is a vector in the general case of true dietary intakes, usual dietary intakes. Z is a vector of adjusting covariates; it could be demographics; it could be anything else. By r I will denote the outcome risk function, and I will explain what I mean by this in a moment. And η denotes a covariate-based predictor where α is a vector of parameters. So we will be mostly considering parametric models.

And so the risk model relates the outcome risk function to the covariate-based predictor.

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)

Slide 10

I will start with examples of the common risk models. It could be linear regression for continuous outcome, logistic regression for binary, or Cox regression for survival analysis.

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.,

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

Slide 11

In the case of linear regression, as I mentioned, Y—outcome—is a continuous variable; for example, blood pressure, cholesterol level, etc. And the risk function in this case is the conditional expectation or conditional mean of this continuous variable given all the covariates in the model.

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.,

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

Slide 12

With the logistic regression, outcome is a binary variable, usually coded as 1 in the case of the event or disease, or 0 otherwise. And risk function is the logit of the probability of event, or odds of event, again conditional on covariates and is given by this expression.

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.,

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

Slide 13

In the case of the Cox regression, outcome is time to event and risk function is log of the hazard function, again conditional on covariates.

Risk models: risk predictor (1)

- A rather 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, $v^* = g(v; \gamma_v)$ denotes its transformed value using Box-Cox family of transformations

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

Slide 14

What about the predictor in the risk model? Well, a rather flexible risk model specifies this predictor as linear over transformed covariates. And we will consider a special class of transformation based on the so-called Box-Cox family of transformations, given by this formula. And, in general, it includes power functions or the logarithm.

Risk models: risk predictor (2)

- 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

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

Slide 15

So again, here is my risk model. And it's important to realize that the slope for a given exposure, T_k let's say, could be interpreted in the following way. Because we allow for transformation of the exposure, due to this fact, the effect of the exposure depends not only on change in this exposure, which would be the case with no transformation of the exposure, but also on its initial value—in other words, effect of change in exposure from some variable T zero by delta T. And the risk is given in this formula; it's the regression coefficient times the difference between the transformed values.

But the bottom line, which is important to emphasize, is that transformation of covariates makes a risk model much more flexible, and yet it doesn't jeopardize the interpretation of the effect on the original scale. In other words, if you consider a power transformation, it will not be interpreting the effect of changing your exposure on the power scale, let's say the square root scale, but on the original scale.

MEASUREMENT ERROR

Slide 16

All right, measurement error….

Dietary measurement error (1)

- Problem in nutritional epidemiology: true usual intakes are **unknown** and measured with error
- Fitting risk models to measured dietary exposures leads to:
  - Bias (often attenuation) of estimated exposure effect
  - Invalid significance tests in models with multiple error-prone covariates due to residual confounding
  - Reduced power to detect exposure effect

Slide 17

A well-known problem in nutritional epidemiology is that true usual intakes are unknown and measured with error, at least in free-living populations. And fitting risk models to measured dietary exposures leads to three unpleasant things: first of all, bias, in most cases attenuation of estimated exposure effect; generally speaking, a possibility of invalid significance tests if the risk model has multiple error-prone covariates, and this is due to residual confounding; and reduced power to detect exposure effect.

Dietary measurement error (2)

- Statistical methods such as regression calibration correct biases due to measurement errors, making statistical tests valid
- These methods do not fully restore the **power** to detect a relationship which is lost due to measurement error
- It is therefore critical to use dietary assessment that, after adjustment for measurement error, leads to the minimum loss of power

Slide 18

There are many statistical methods such as regression calibration, on which we will concentrate in this lecture, that correct biases due to measurement error, also making statistical tests valid. It's important to realize that those methods do not fully restore the power to detect a relationship which is lost due to measurement error. And I will talk about it in some more detail in a moment. But the important thing here is to emphasize that because of this loss of power it's critical to use dietary assessment that, after adjustment for measurement error, leads to the minimum loss of power.

REGRESSION CALIBRATION

Slide 19

All right, so as I said, we will concentrate on regression calibration in this talk.

Regression calibration (1)

- Denote measured dietary intake by $\mathbf{D} = (D_1, \ldots, D_p)^t$
- Assumption: measurement error in **D** is **non-differential** with respect to health outcome *Y*, i.e., provides no additional information about *Y* 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

Slide 20

Let's denote the observed or measured dietary intake by D; this is a vector with k components. The main assumption that we're going to use is that measurement error in D is nondifferential with respect to health outcome, Y, which means that it provides no additional information about Y beyond that in true diet. This assumption may be justified in cohort studies because diet there is usually assessed before outcome is known, usually at the baseline, but not necessarily in case-control studies due to possible recall bias when cases report their past diet differently from noncases. And this is the main reason, actually, why I will concentrate on cohort studies.

Regression calibration (2)

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

$$T_k^{*P}(\mathbf{D},\mathbf{Z}) = \mathbb{E}(T_k^* \mid \mathbf{D},\mathbf{Z}),\ k = 1,\ldots,p,$$

  i.e., its conditional mean (expectation) given all measured dietary components **D** and error-free covariates **Z** in the risk model

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

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

Slide 21

The regression calibration consists of substituting for each error-prone covariate in the risk model its best predictor, or its conditional mean—or as statisticians like to call it, conditional expectation—given everything that has been observed, which means the observed diet, D, and error-free covariates, Z, in the model. And if one does this, it leads to approximately the same regression slopes as if one would use the true dietary intakes. The only exception is the intercept, but we don't care much about it.

Regression calibration (3)

- The precision of estimated slopes α_{T_k} in risk model

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

and the power to detect dietary effects depend not on dietary data **D** themselves, but on the precision of calibration predictors $T_k^{*P}(\mathbf{D}, \mathbf{Z}) = \mathbb{E}(T_{ki}^* \mid \mathbf{D}_i, \mathbf{Z}_i)$ to predict true transformed intakes $T_{ki}^*,\ k = 1, \ldots, p.$

Slide 22

Now, the precision of the estimated slope after the substitution, and therefore the power to detect dietary effects, depends not on the original data, D, themselves but on the precision of the calibration predictor to predict true transformed intakes. What it basically means is that although we use the same information, which comes from observed intakes, D, and covariates, Z, the information is packed or used in slightly different ways. We don't use a particular component of D for this particular intake of interest, but we use the predictor of true intake given all the observed intakes and all of the covariates.

Regression calibration (4)

- For practical reasons (relatively low cost and possible mass mailings), assessment of diet in nutritional epidemiology has been commonly done by food frequency questionnaires (FFQs)
- As you learned in webinar 10, repeat administration of more precise short-term instruments, such as 24-hour dietary recall (24HR) or food records (FR), may substantially improve the precision of the calibration predictor and the power to detect a dietary effect

Slide 23

For practical reasons, assessment of diet in nutritional epidemiology has been commonly done by food frequency questionnaire (FFQ), and the reasons are pretty obvious: relatively low cost and the possibility of mass mailings.

You all learned in webinar 10 given my Doug Midthune that repeat administrations of more precise short-term instruments such as 24 hour dietary recalls—I will call them 24HR—or food records may substantially improve the precision of the calibration predictor and, therefore, increase the power to detect a dietary effect.

Regression calibration (5)

- Until recently, repeat application of short-term instruments as the main dietary assessment method in large studies was prohibited by a high cost of their administration and/or processing
- With advancement of new technology, repeat administration of much less expensive automated short-term instruments (e.g., web-based ASA24 developed at NCI) has become a reality

Slide 24

Until recently, repeat application of short-term instruments as the main dietary assessment method in large studies was prohibited by the high cost of their administration or of their processing. There are a few exceptions. There are several cohorts that I know of in the UK where the main instrument was the repeated seven-day food record. But I must say that the full data set has not been processed up until now. So they expect to do it, I think, in a year or so.

But with advancement of new technology, repeat administration of much less expensive automated short-term instruments—an example would be a Web-based 24 hour recall developed at NCI which we call ASA24—now is a reality.

Regression calibration (6)

- Additional way to improve the precision of the calibration predictor is to consider **enhanced regression calibration:**
  - Let vector **X** include error-free covariates **Z** in the risk model and additional covariates **C** that are related to true intakes but not to outcome given true intakes
  - Predictor $\mathbb{E}\left(T_{ki}^{*} \mid \mathbf{D}_i, \mathbf{X}_i\right)$ is not only legitimate to use in regression calibration but is generally more precise than predictor $\mathbb{E}\left(T_{ki}^{*} \mid \mathbf{D}_i, \mathbf{Z}_i\right)$

Slide 25

Before considering this design with the main instrument being a short-term unbiased measure, let us concentrate on yet another way to improve the precision of the calibration predictor. It's called enhanced regression calibration. Consider vector, X, which includes all error-free covariates, Z, in the risk model and additional covariates—I call them C—that are related to true intakes but not to outcome given true intakes. In other words, covariates in C are not confounders but they may help in predicting true intakes.

And so instead of considering the predictor of true intake given D and Z, now we consider the predictor of true intake given D and X. And this predictor is not only legitimate to use in the regression calibration but, generally speaking, it provides more precision than the usual regression calibration predictor.

Regression calibration (7)

- In what follows we will consider regression calibration when dietary assessment is done with repeat short-term measurements $\mathbf{R}_{ki} = \left(R_{ki1}, \ldots, R_{kiJ_i} \right)^t$
- Ideally, when FFQ is also administered, it will be used in enhanced RC as part of vector **C**
- With advancement of new technology, cohort studies with repeat automated short-term instruments, alone or in combination with FFQ, are now being planned
- In what follows, we present a newly developed methodology for correcting results of such studies for measurement error

Slide 26

So in what follows we will consider regression calibration when dietary assessment is done with repeat short-term unbiased measurements. I will call them R. And, ideally, we would like to use enhanced regression calibration, and a good example of this component of vector C is the FFQ. And so we will consider designs where, ideally, the main instrument consists not only of the repeat short-term measurements, R, but also includes the FFQ in the main study.

With the advancement in new technology, as I mentioned, cohort studies with this design are now being planned, in many countries actually. And so in anticipation of such studies, we present here a newly developed methodology for correcting results of these studies for measurement error. And, by the way, it may help in analyzing the existing cohorts in the UK, as I mentioned before.

Regression calibration (8)

- Main assumption: for person *i*, repeat *j*, short-term measurement R_{kij} is unbiased for true usual intake
- Regression calibration predictor is given by

$$T_{ki}^{*P} = \mathbb{E}\left[g_{T_k}\left(T_{ki}\right) | \mathbf{R}_i, \mathbf{X}_i \right] = \mathbb{E}\left[g_{T_k}\left\{ \mathbb{E}\left(R_{kij} | i\right)\right\} | \mathbf{R}_i, \mathbf{X}_i \right]$$

- Short of averaging an infinite number of repeat measurements, evaluation of conditional means in the above formula requires modeling of R_{kij}
- Having a model, expectations can be evaluated as integrals over corresponding distributions

Slide 27

So our main assumption is that for person i and repeat measure j, short-term measurement R_{kij} on the k^{th} dietary component is unbiased for true usual intake. And in this case, the regression calibration predictor is given by this formula. Okay, so remember, we are considering transformed intake, g(T), and T is the conditional expectation of R, due to our assumption, given all personal information, and we have to calculate its conditional mean given what is observed, R and X. And, right here, you can note the difference between this regression calibration and the regression calibration that we considered before. What we considered before was we regressed a reference instrument on FFQ. Here, we use the short-term instrument to measure the true intake and also as a covariate in the regression calibration.

Now, if you look at this expectation—for example, this expectation inside expectation—in principle it could be done by averaging an infinite number of repeat measurements for each individual. But of course we don't have this infinite number of repeat measurements. And so to evaluate this conditional mean, this formula requires a model for R_{ij}. Having a model, the expectations could be evaluated as integrals because in statistical terms expectation is an integral over the distribution of the involved random variables.

Regression calibration (9)

- Methodology below is developed for any unbiased repeat short-term measurements
- This methodology is demonstrated using 24HR
- **Working assumption**: 24HR is unbiased in reporting individual's true usual dietary intake
  - Implications of possible biases in 24HR are discussed at the end

Slide 28

The methodology that we have developed could be applied to any unbiased repeat short-term measurements. I will demonstrate this methodology using the 24HR. So the working assumption throughout this lecture is that the 24HR is unbiased in reporting individuals' true usual dietary intake. And at the end I will discuss the implications of possible biases in 24 hour recalls.

SINGLE DIETARY EXPOSURE

Slide 29

I will start with the single dietary exposure. It’s easier to explain all the involved concepts in this particular case, although of course such models are not very realistic in practice, where risk models usually have several dietary exposures. But we will go to it, too.

Regularly-consumed dietary components (1)

- **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 is thus specified as

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

Slide 30

I will start with the ideal world of the classical measurement error model, which means that the observed measurements, R, are truth plus error. This error is independent of true intake. It has a normal distribution with mean zero and a constant variance. And I will also assume that the true intake being regressed on vector X has linear regression. In other words, it is given by this formula where the residual from this regression has a normal distribution with mean of zero and a constant variance.

And putting those two formulas together, the measurement error model is specified as the following expression.

Regularly-consumed dietary components (2)

- 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

Slide 31

Now let's look a little bit more carefully at this expression. It's what is called in statistics a mixed effects linear model. It includes fixed effects. In this case they are linear effects of covariates, X, defined by the population-level parameters; so β_0, β_x, are the parameters, which are the same for any individual in the population. It also contains the random effect, ui, representing part of the within-person mean that is not explained by the covariates. So two persons may have the same covariates but their reported intake would be different because this random effect, u, is different for those two persons. So this effect is person-specific but randomly varies across people. And, of course, like in other linear models, we have a within-person random error, ε, representing short-term variation.

Regularly-consumed dietary components (3)

- Let $\boldsymbol{\theta}$ denote parameters of the measurement error model for **R**
- RC predictor of transformed true usual intake is given by

$$T_i^{*P} = \mathbb{E}\left[T_i^* \mid \mathbf{R}_i, \mathbf{X}_i\right] = \int g_T\left(\beta_0 + \boldsymbol{\beta}_X^t \mathbf{X}_i + u_i\right) f\left(u_i \mid \mathbf{R}_i, \mathbf{X}_i; \boldsymbol{\theta}\right) du$$

- Evaluation of this expression requires evaluation of the probability density function (pdf)

$$f\left(u_i \mid \mathbf{R}_i, \mathbf{X}_i; \boldsymbol{\theta}\right)$$

which defines the conditional distribution of *u*

Slide 32

Let vector θ denote parameters in the measurement error model for R. So the regression calibration predictor of transformed true usual intake is given by this formula. Remember, I said if we have a model, the conditional mean is just an integral. Here it is an integral of this function. It's a transformed true intake, which depends on X and u, and given R and X the only random variable that could vary is u. So we need to take this integral over the distribution of u given R, given X, and given the parameters in the measurement error model.

So to evaluate this integral, one requires the knowledge of this distribution, called the probability density function or pdf.

Regularly-consumed dietary components (4)

- According to Bayes' theorem

$$f\left(u_i \mid \mathbf{R}_i, \mathbf{X}_i; \theta\right) = \frac{f\left(\mathbf{R}_i \mid \mathbf{X}_i, u_i; \theta\right) f\left(u_i \mid \mathbf{X}_i; \theta\right)}{\int f\left(\mathbf{R}_i \mid \mathbf{X}_i, u_i; \theta\right) f\left(u_i \mid \mathbf{X}_i; \theta\right) du}$$

where, given parameters θ, conditional pdf's on the right are defined by the distributions of ε_{ij} and u_i

- When θ is estimated by fitting the measurement error model to data, $\hat{T}_i^{*P}$ is known as the Empirical Bayes's (EB) estimator

Slide 33

And we can evaluate this density function based on this formula, which is based on the well-known Bayes' theorem. So if you look at the numerator, it has a product of two pdf's and given parameters, θ, and given the X in this case and X and u in this case, those pdf's are fully defined by the distribution of ε and u, which in our measurement error model are specified as normal distributions.

So to calculate this expression, the only thing we need to do is to fit the measurement error model to estimate parameters θ. And so when parameters θ are estimated by fitting the measurement error model to available data, the regression calibration predictor, or estimated regression calibration predictor, is known as the Empirical Bayes' Estimator.

Regularly-consumed dietary components (5)

- If the risk model is fitted on original scale ($g_T(T) \equiv T$) RC predictor exists in closed form and is known as the Best Linear Unbiased Predictor (BLUP) given by

$$\hat{T}_i^P = \hat{w}\bar{R}_i + (1-\hat{w})\left(\hat{\beta}_0 + \hat{\boldsymbol{\beta}}_X^t \mathbf{X}_i\right)$$

where

$$\hat{w} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_\varepsilon^2 / J_i}, \quad \bar{R}_i = \frac{1}{J_i}\sum_{j=1}^{J_i} R_{ij}$$

- In general, RC predictor $\hat{T}_i^{*P}$ of transformed intake is not linear, does not exist in closed form, and has to be evaluated by numerical integration

Slide 34

Consider the simplest case when the risk model is fitted on the original scale; in other words, the transformation, g, is identity. In this case, the regression calibration predictor exists in closed form and is well known as the Best Linear Unbiased Predictor, or BLUP, and is given by this formula. It's the weighted average between the fixed effects and the individual mean, where weights are given by this formula.

In general, though, when g is not an identity transformation, the regression calibration predictor of transformed intake is not linear, so it's not given by BLUP and does not exist in closed form, and one has to calculate the corresponding integrals using numerical integration.

Regularly-consumed dietary components (6)

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

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

where:

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

Slide 35

That was an ideal world. Unfortunately, or maybe fortunately, we all live in the real world and in the real world, often, within-person random error in R depends on true intake and has a skewed distribution, violating classical model assumptions. The usual remedy in this case is to transform R to a scale where the classical model is a good approximation. So after this transformation we assume that transformed R is the sum of within-person mean—I call it μ—and within-person variation—I call it ε; ε has a normal distribution with mean zero and a constant variance; and μ has a linear regression on X with the residual of this regression again being normally distributed with mean zero and constant variance.

Regularly-consumed dietary components (7)

- On the transformed scale we have

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

- Measurement error model is then specified as non-linear mixed effects model

$$R_{ij}=g_R^{-1}\left(\beta_0+\boldsymbol{\beta}_X^t\mathbf{X}_i+u_i+\varepsilon_{ij}\right)$$

- Denoting by θ model parameters, R_{ij} is the function

$$R_{ij}=\Re(\mathbf{X}_i,u_i,\varepsilon_{ij};\boldsymbol{\theta})$$

Slide 36

In this case, on the transformed scale, the measurement error model is given by our old friend. It's a mixed effects linear model. But on the original scale, the measurement error model is a nonlinear mixed effects model given by this formula. We have to transform back to the original scale, and this is an inverse transformation.

What's important to realize is that on the original scale, R is the function of X, u, ε, and of course parameters in the measurement error model.

Regularly-consumed dietary components (8)

- Since R_{ij} is unbiased for true intake on original scale

$$T_i = \mathbb{E}\left(R_{ij} \mid i\right) \equiv \mathbb{E}\left\{\Re\left(\mathbf{X}_i, u_i, \varepsilon_{ij}; \boldsymbol{\theta}\right) \mid X_i, u_i\right\}$$

or

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

- RC predictor of transformed intake is EB estimator

$$\hat{T}_i^{*P} = \int g_T\left(\mathfrak{T}(\mathbf{X}_i, \mathbf{u}_i; \hat{\boldsymbol{\theta}})\right) f\left(u_i \mid \mathbf{R}_i, \mathbf{X}_i; \hat{\boldsymbol{\theta}}\right) du$$

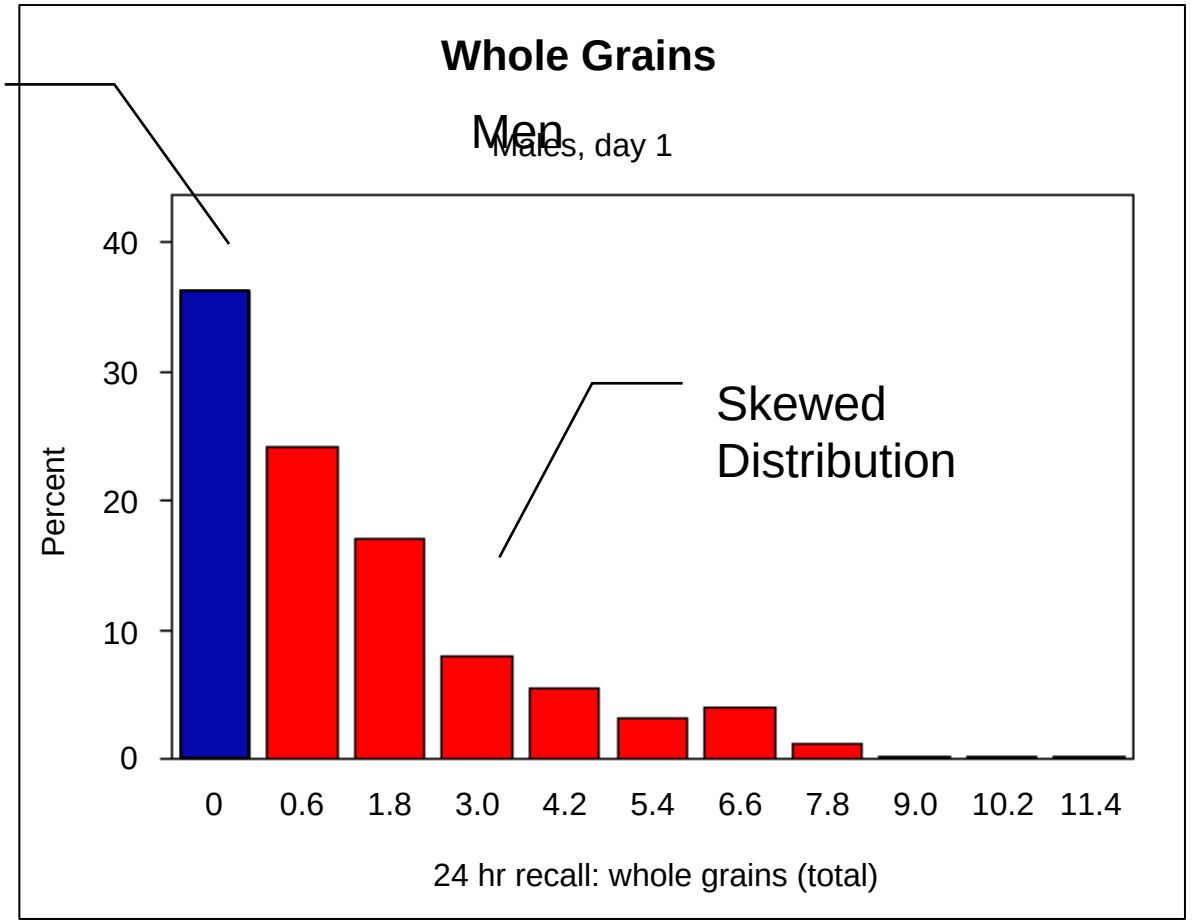
where, as before, pdf $f\left(u_i \mid \mathbf{R}_i, \mathbf{X}_i; \hat{\boldsymbol{\theta}}\right)$ is evaluated using Bayes' theorem

Slide 37

And so since R is assumed to be unbiased for true usual intake on the original scale—remember, the true usual intake is given as its expectation given all personal information, so it's the expectation of this function. And the personal information here consists of vector X and person-specific random effects, u. So to calculate this conditional mean, we have to take an integral of this function over the distribution of ε. And it's given by this formula. And this integral is nothing else than another function, which does not include ε anymore; ε was integrated out, so it's just a function of X, u, and the vector of parameters. And so the regression calibration predictor is the conditional expectation of this function, T, over the distribution of u given R and X. The same way as it was before, again, if the parameters of the measurement error model are estimated by fitting the model to the data, this function is evaluated using the Bayes' theorem and the regression calibration predictor is the Empirical Bayes' Estimator.

Episodic dietary components (1)

- Consider now **episodically-consumed dietary components**



Slide 38

Now, so far, I have considered, although the real world, but the world where R, our measurements, were continuous, which means that the consumption took place every day or almost every day. Let me now switch gears and consider episodic dietary components.

And here is an example; you've seen it several times. It's a histogram of the whole grains in the EATS study reported by men on one 24 hour recall. As you may see, about 36 or 37 percent of men reported zero intake of whole grains on that 24 hour recall, and the rest reported positive intake, which has a skewed distribution with an ugly long right-hand tail.

Episodic dietary components (2)

- Short-term measurements for an episodically-consumed dietary component is a **semicontinuous variable** with **excess zeros** and often **skewed** to the right positive values
- Measurement error model is the result of **two** distinct, although generally correlated **processes**:
  - One specifies binary indicator variable of short-term consumption

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

  - Other specifies its positive value

Slide 39

So short-term measurements for an episodically consumed dietary component are semicontinuous variables with excess zeros and often skewed to the right positive values. And the measurement error model, if you will recall, is the result of two distinct, although generally correlated processes. The first one specifies the binary indicator variable of short-term consumption. So if the reported intake is greater than zero, this binary indicator variable takes a value of 1; otherwise, it is 0.

And the other process specifies the positive value of the consumption.

Episodic dietary components (3)

- **Part I**: Modeling binary indicator of consumption
- Based on modified NCI method (webinar 8), consider continuous **latent variable**

$$\tilde{R}_{ij} = \beta_{10} + \boldsymbol{\beta}_{1X}^{t}\mathbf{X}_{i} + u_{1i} + \varepsilon_{1ij},\ u_{1i} \sim N\left(0, \sigma_{u_1}^{2}\right),\ \varepsilon_{1ij} \sim N(0,1)$$

which underlies fact of consumption in period j

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

- Consumption probability is given by the probit model

$$\mathbb{P}(R_{ij} > 0 \mid i) = \Phi\left(\beta_{10} + \boldsymbol{\beta}_{1X}^{t}\mathbf{X}_{i} + u_{F1i}\right)$$

where Φ denotes standard normal distribution function

Slide 40

So we're dealing with a two-part model. Part I is modeling binary indicator of consumption. It could be specified in two ways. You can either specify the probability of this indicator variable to take the value of 1; in other words, the probability to have consumption on any given day. And this was done in the original NCI method. Or you can try and model the fact of consumption. And this is a modified NCI method, which I discussed in webinar 8. And so we will go with this one here, so consider a latent variable, R tilde, given by this linear mixed effects model where u has a normal distribution with a constant variance and ε has a normal distribution with variance of 1. The reason I put 1 here is the model identifiability because the fixed effects of this model are identified or could be estimated uniquely only proportional to the variance of ε. So we may as well make this variance be equal to 1.

And I will make this latent variable underlie the fact of consumption in period j; in other words, I will say that consumption takes place if and only if this latent variable is positive. And in this case, by the way, if one wants to calculate the consumption probability, it's given by the mixed effects probit model, given by this formula.

Episodic dietary components (4)

- **Part II**: Modeling amount during consumption period
- Given consumption in period j, transformed amount is specified as mixed effects linear model

$$g_R(R_{ij} \mid R_{ij} > 0) = \beta_{20} + \boldsymbol{\beta}_{2X}^t \mathbf{X}_i + u_{2i} + \varepsilon_{2ij}$$

where

$$u_{2i} \sim N\left(0, \sigma_{u_2}^2\right), \ \varepsilon_{2ij} \sim N\left(0, \sigma_{\varepsilon_2}^2\right)$$

- Person-specific random effects u_{1i}, u_{2i} in parts I and II are allowed to be correlated to induce correlation between probability to consume and consumption amount

Slide 41

Part II, modeling amount during consumption period—this is our old friend. On the transformed scale, this is the mixed effects linear model. The only thing here which is important to emphasize is that we get to observe these distributions of u and ε only when consumption takes place. And of course to make the probability to consume and the consumption amount to be correlated, we allow random effects, u_1 and u_2, in both parts of the model to be correlated.

Episodic dietary components (5)

- Measurement error model is formally specified as non-linear mixed effects model

$$R_{ij} = I\left(\beta_{10} + \boldsymbol{\beta}_{1X}^{t}\mathbf{X}_i + u_{1i} + \varepsilon_{1ij} > 0\right) g_R^{-1}\left(\beta_{20} + \boldsymbol{\beta}_{2X}^{t}\mathbf{X}_i + u_{2i} + \varepsilon_{2ij}\right)$$

where

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

$$\boldsymbol{\varepsilon}_{ij} = (\varepsilon_{1ij}, \varepsilon_{2ij})^t \sim N\left\{\mathbf{0}; \begin{pmatrix} 1 & 0 \\ 0 & \sigma_{\varepsilon 2}^2 \end{pmatrix}\right\}$$

Slide 42

So, formally, the measurement error model is specified as given by this formula. So this is the indicator variable of the fact of consumption. Remember, this is a latent variable. If it's positive, consumption takes place. And so this indicator variable equals to 1 if consumption takes place. And then the consumption amount on the original scale is given by the inverse transformation of this expression. If this is not true, if this latent variable is not positive, this is zero and of course we observe zero consumption.

So now u is a vector; u_1 and u_2 are allowed to be correlated with unrestricted correlation parameters. As to the vector, ε, ε_1 has variance of 1. Since the distribution of ε_2 is defined given the fact of consumption, it's specified as uncorrelated with ε_1. So this covariance is 0.

Episodic dietary components (6)

- Denoting by $\boldsymbol{\theta}$ model parameters, we have:

$$R_{ij} = \Re(\mathbf{X}_i, \mathbf{u}_i, \boldsymbol{\varepsilon}_{ij}; \boldsymbol{\theta})$$

- True usual intake of episodic component is given by

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

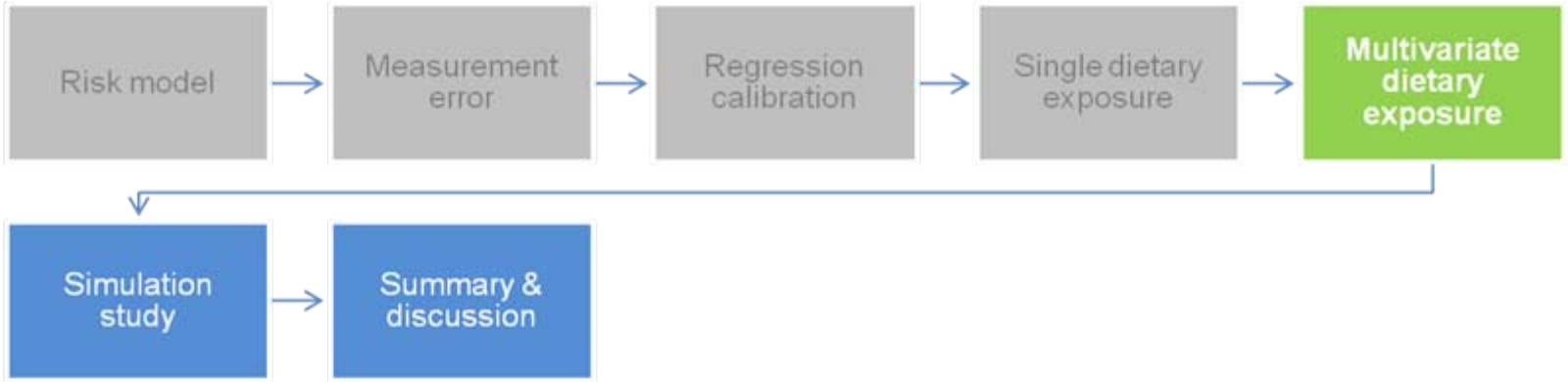
- RC predictor of transformed intake is EB estimator

$$\hat{T}_i^{*P} = \int g_T\left(\mathfrak{T}(\mathbf{X}_i, \mathbf{u}_i; \hat{\boldsymbol{\theta}})\right) f\left(\mathbf{u}_i \mid \mathbf{R}_i, \mathbf{X}_i; \hat{\boldsymbol{\theta}}\right) d\mathbf{u}$$

where, as before, pdf $f\left(\mathbf{u}_i \mid \mathbf{R}_i, \mathbf{X}_i; \hat{\boldsymbol{\theta}}\right)$ is evaluated using Bayes' theorem

Slide 43

Again denoting by θ model parameters, our measurement, R, is a function of X, u, and ε. The true intake is the expectation of this function given X and u, or the integral over the distribution of ε given X and u. It's another function. We integrated out ε so it's a function of X and u, and of course the vector of parameters. And estimating the parameters by fitting the measurement error model to the data, the regression calibration predictor is an Empirical Bayes' Estimator given by this expression, where, as before, this probability density function is estimated using the Bayes' theorem. And integration in this general case is done numerically.



MULTIVARIATE DIETARY EXPOSURE

Slide 44

This was the univariate case to describe the main concepts. And, by the way, in this univariate case, this description was published in our *Biometrics* paper in 2009.

And let me now go to the multivariate dietary exposure, which is much more realistic. And this is completely new; this development is new and has never been published. So this is the first time.

Multivariate measurement error model (1)

- Consider risk models with *p* dietary components, the first *m* of which are episodically-consumed and last *p - m* are daily consumed
- Multivariate measurement error model is specified as

$$
\begin{aligned}
R_{kij} &= I\left(\beta_{2k-1,0} + \boldsymbol{\beta}^{t}_{2k-1,X}\mathbf{X}_i + u_{2k-1,i} + \varepsilon_{2k-1,ij} > 0\right) \\
&\quad \times g^{-1}_{R_k}\left(\beta_{2k,0} + \boldsymbol{\beta}^{t}_{2k,X}\mathbf{X}_i + u_{2k,i} + \varepsilon_{2k,ij}\right),\ k = 1,\ldots,m; \\
R_{kij} &= g^{-1}_{R_k}\left(\beta_{m+k,0} + \boldsymbol{\beta}^{t}_{m+k,X}\mathbf{X}_i + u_{m+k,i} + \varepsilon_{m+k,ij}\right),\ k = m+1,\ldots,p
\end{aligned}
$$

Slide 45

So consider risk models with p dietary components. I will consider the case when m out of p are episodically consumed and the rest of them are daily consumed. The multivariate measurement error model is specified basically as before, so in the case of the episodically consumed dietary component R is this product of the indicator variable of consumption and inverse transformation of this function. And in the case of the daily consumed dietary components, of course consumption always takes place, the probability of consumption is 1, so R is just given by this function.

What's important here is to specify the multivariate distribution of u's and the multivariate distribution of ε's. So the vector u, which is now a long vector of m plus p components, follows a normal distribution with the mean being the vector of zeros and unrestricted variance-covariance matrix. Vector ε has a normal distribution with mean zero and a variance-covariance matrix which is a little bit more peculiar. It consists of four parts. Those three submatrices, there, are allowed to roam the way they wish, but this submatrix is a special submatrix. It's a patterned submatrix; it's a structured submatrix. Because, remember, when we considered one episodic dietary component, it had the following form: 1 in the upper left-hand corner, parameter sigma 2 in the lower right-hand corner, and two zeros. Now, those blocks are going to be diagonal blocks for all m episodic dietary components. Those other elements are arbitrary.

Multivariate measurement error model (2)

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

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

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

$$\boldsymbol{\Sigma}_{\varepsilon} = \begin{pmatrix} \boldsymbol{\Sigma}_{\varepsilon 11} & \boldsymbol{\Sigma}_{\varepsilon 12} \\ \boldsymbol{\Sigma}_{\varepsilon 21} & \boldsymbol{\Sigma}_{\varepsilon 22} \end{pmatrix}, \quad \boldsymbol{\Sigma}_{\varepsilon 11} = \begin{pmatrix} 1 & 0 & \sigma_{13} & \ldots & \sigma_{1,2m} \\ 0 & \sigma_2^2 & \sigma_{23} & \ldots & \sigma_{2,2m} \\ \ldots & & & \ldots & \\ & & & 1 & 0 \\ & & & 0 & \sigma_{2m}^2 \end{pmatrix}$$

Slide 46

Now what does that mean? First of all, let's start with the variance-covariance matrix of u's. We allow all possible correlations among those components, which would induce all kinds of correlations.

Multivariate measurement error model (3)

- Model characteristics:
  - Allowing correlations among all person-specific random effects

$$\boldsymbol{\Sigma}_{\mathbf{u}} = \begin{pmatrix} \sigma_{u11} & \cdots & \sigma_{u1,m+p} \\ \vdots & \ddots & \vdots \\ \sigma_{u1,m+p} & \cdots & \sigma_{u,m+p,m+p} \end{pmatrix}$$

induces correlations among usual intakes of regular and episodic components (within each group and across two groups)

Slide 47

Thus probabilities of consumption of episodic components and usual amounts of episodic components, if intake takes place, are allowed to be correlated. Usual amounts of daily consumed dietary components are allowed to be correlated. And the usual amounts of episodic components if greater than zero are allowed to be correlated with the usual amounts of the daily components.

Multivariate measurement error model (4)

- Model characteristics (continuation): allowing correlations among within-person errors induces
  - Correlation among intakes of regular and episodic components (within each group and across two groups) during a short-term consumption period
  - Correlations among indicators of short-term consumption for different episodic components
  - Correlations among an indicator of consumption for any episodic component and intakes of regular components during a short-term period

Slide 48

Now, what about the correlations imposed by this patterned matrix, variance-covariance matrix, of ε's? This pattern allows the following things: first of all, correlations among intakes of regular and episodic components within each group and across the groups during a short-term period of consumption, so for a 24 hour recall on each day. And the examples in this case could be—let me start with the correlation among daily components which are daily consumed; for example, fat and protein intake, could be correlated.

Now, it also allows correlations among indicators of short-term consumption for different episodic components. For example, let's say that on a particular day red meat consumption did take place. It may mean that on that day there will not be consumption of fish.

Then, we allow correlations among an indicator of consumption for any episodic component and intakes of daily or regular components during any short-term period. For example, the fact that red meat was consumed on a particular day may mean that intake of fat and protein or energy will go up on that day.

Multivariate measurement error model (5)

- Denoting by θ model parameters, we have:

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

- Multivariate true usual intake is given by

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

- RC predictor of transformed intake is EB estimator

$$\hat{\mathbf{T}}_i^{*P} = \int \mathbf{g}_T\left(\mathfrak{T}(\mathbf{X}_i,\mathbf{u}_i;\hat{\boldsymbol{\theta}})\right) f\left(\mathbf{u}_i \mid \mathbf{R}_i,\mathbf{X}_i;\hat{\boldsymbol{\theta}}\right) d\mathbf{u}$$

where, as before, pdf $f\left(\mathbf{u}_i \mid \mathbf{R}_i,\mathbf{X}_i;\hat{\boldsymbol{\theta}}\right)$ is evaluated using Bayes' theorem

Slide 49

Denoting model parameters by θ, R again is a function of X, u, and ε, much more complicated because u and ε are now vectors of many, many components, but it is still a function. And multivariate true usual intake is given as an integral of this function over the distribution, multivariate distribution, of ε given X and u. So we're dealing with the multivariate function of X and u and model parameters, and the regression calibration predictor—and it's a vector—is the integral of the transformation of that function over the multivariate distribution of u given R and X.

And this probability density function, multivariate probability density function, again could be evaluated using the Bayes' theorem and then the RC predictor is an Empirical Bayes' Estimator.

What's important to realize here is that we're dealing with the multivariate distribution, and because of this, in this particular case the regression calibration could not be done one variable at a time. It has to be done for all involved dietary components. It has to be done simultaneously.

Multivariate measurement error model (6)

- New multivariate measurement error model is a highly non-linear mixed effects model with many correlated latent variables and patterned covariance matrix Σ_ε with structured zeros and ones
- Currently available software for MLE or EM fitting cannot handle such models
- The model is therefore fitted using Markov Chain Monte Carlo paradigm
- Working version of SAS program has been developed by Dennis Buckman
- A user-friendly version is under construction

Slide 50

So this new multivariate measurement error model is a highly nonlinear mixed effects model with many correlated latent variables and a patterned or structured covariance matrix of ε's with structured zeros and ones. And the problem with this is that, currently, no available software for maximum likelihood estimation or EM estimation, which is the expectation minimization approach, would be able to fit such a model. And so we therefore used the Markov Chain Monte Carlo paradigm, which you should be a little bit familiar with from the webinar that Raymond Carroll gave.

The working version of this Markov Chain Monte Carlo program has been developed by Dennis Buckman, and he is working now on a user-friendly version. It's under construction and once it is ready it will be on our Web site.

SIMULATION STUDY

Slide 51

All right, simulation study ….

Simulation study (1)

- Data: generated FFQ and 1000 24HRs for 2000 subjects, with distributions similar to those of red meat, white meat, total fruit, and energy in NIH-AARP calibration study of men
- True usual intakes: calculated as averages of 1000 24HRs; density intakes were calculated as ratios of true usual components to usual energy intakes
- Binary disease outcome (e.g., cancer or no cancer): generated using probability of disease defined by logistic regression based on specified odds ratios (OR) for each of the 3 true densities and energy

Slide 52

As I mentioned before, this new methodology has been developed in anticipation of the future studies with automated short-term instruments as the main instrument of the study. So at the moment we actually don't have any real data to fit the model and show how it works. Therefore, we decided to do a simulation study, and a big plus of a simulation study is that when you simulate something you know the truth. So we can apply the new methodology and check whether it will allow us to recover the true parameters in the risk model.

So we generated the data. We generated the FFQ and 1,000 24 hour recalls for 2,000 subjects with distributions similar to those of red meat, white meat, total fruit, and energy—four dietary components—in the NIH-AARP calibration study of men.

So remember, NIH-AARP is a large cohort with a calibration substudy with about 1,000 men and 1,000 women. We considered men here. And so in this calibration substudy we have data on an FFQ and two 24 hour recalls and that's how we modeled the distribution and generated our data.

The true usual intakes were calculated as averages of those 1,000 24 hour recalls for each subject. And then we took densities, which we calculated as the ratios of true usual components—in this case, white meat, red meat, and total fruit—to usual energy intake.

We've also simulated the binary disease outcome. You can think of it as cancer or no cancer, disease or no disease. We generated it using the probability of disease as defined by the logistic regression based on prespecified odds ratios for each of the three true densities and true energy intake.

Simulation study (2)

- Simulated cohort: 2000 subjects with 2 24HRs (took first 2 of 1000 simulated), FFQ and binary disease outcome
- Goal: estimating log OR of disease for increasing:
  - Red meat between 10 & 60 g/1000 kcal
  - White meat between 10 & 60 g/1000 kcal
  - Total fruit between 0.2 & 1.0 cups/1000 kcal
  - Total energy between 1500 and 3000 kcal
- Risk model: logistic regression with standard errors estimated by bootstrap

Slide 53

After that, we simulated the cohort, which consisted of 2,000 subjects with two 24 hour recalls—so we took the first 2 out of 1,000 that we simulated—and the FFQ and of course the binary disease outcome that we simulated. That was our cohort.

Our goal was to estimate the log odds ratio of disease for all four dietary components: red meat density between 10 and 60 g/1,000 kcal; white meat density between 10 and 60 g/1,000 kcal, total fruit between 0.2 and 1 cups/1000 kcal, and total energy between 1500 and 3000 kcal. And our risk model was logistic regression.

Simulation study (3)

Reported consumption of red meat, white meat, total fruit, and total energy on two 24HRs

	Red meat (g/day)	White meat (g/day)	Total fruit (cups/day)	Total energy (kcal/day)
Mean intake (s.e.)	82.2 (0.12)	75.5 (0.12)	1.66 (0.002)	2299.6 (1.11)
Mean amount on consumption days (s.e.)	112.1 (0.13)	116.3 (0.15)	1.77 (0.002)	2299.6 (1.11)
Probability to consume	0.75	0.67	0.92	1
% who consumed:				
0 out of 2 days	14.01	18.38	2.54	0
1 out of 2 days	29.77	36.37	12.63	0
2 out of 2 days	56.22	45.25	84.83	100

Slide 54

Here are some characteristics of the simulated cohort. The first row gives you the mean of two 24 hour recalls for all four dietary components and the standard error. The next row gives the means, the same means, but on consumption days only. So red meat is basically an episodically consumed dietary component, so you may see that the mean on consumption days is larger than the mean overall because this one includes days with zero intakes; the same for white meat. For total fruit, the difference is not that large because total fruit is somewhat episodic but not much. And total energy is always consumed and so those two figures are exactly the same.

This is the probability to consume. Now, remember that the probability to consume differs for different subjects. It's a function of covariates; it's a function of person-specific random effect. So this is just the mean. For red meat it is .75; for white meat it's a little bit smaller. For total fruit it's not quite 1 but close, .92. It's 1 for total energy.

Here, we have the percentage of subjects in the cohort who didn't report consumption on any of the two 24 hour recalls. For red meat 14 percent didn't; for white meat a little bit more than 18 percent had both zeros. For total fruit it's only 2.5 percent, and of course for energy it's zero. Those are both percentages for those who reported consumption one out of two days, and the last row displays the percentages of subjects who reported consumption on both days. And as you may see, it's a little bit more than 50 percent for red meat. It's a little bit less than 50 percent for white meat. It's about 85 percent for total fruit, and 100 for energy.

Simulation study (4)

Mean and standard deviation of estimated log odds ratio in logistic regression of disease on red meat, white meat, total fruit, and energy

Dietary exposure	Covariates in risk model	Mean Log OR (s.e.)	
Red meat density	**True intakes**	**0.4**	
10 – 60 g/1000 kcal	Mean 24HR	0.142	(0.008)
	Enhanced RC predictor	0.395	(0.023)
White meat density	**True intakes**	**0**	
10 – 60 g/1000 kcal	Mean 24HR	-0.057	(0.008)
	Enhanced RC predictor	0.006	(0.023)
Total fruit density	**True intakes**	**-0.2**	
0.2 – 1 cups/1000 kcal	Mean 24HR	-0.155	(0.007)
	Enhanced RC predictor	-0.223	(0.012)
Total energy kcal	**True intakes**	**0.2**	
1500 – 3000 kcal	Mean 24HR	0.076	(0.011)
	Enhanced RC predictor	0.224	(0.021)

Slide 55

This table gives the results of fitting the model, and we compared fitting the models to mean 24 hour recall—so this we call the naïve approach—and also to predictors using the enhanced regression calibration. I say enhanced because, remember, we simulated FFQ in addition to 24 hour recall. This FFQ was used as a covariate in the measurement error model, so this is a component of vector C, the additional vector that we considered—therefore, the enhanced regression calibration predictor.

First of all, the truth that we used to simulate the outcome: For red meat density for this increase from 10 to 60 g/1000kcal the true log odds ratio was .4, so it's a risk factor. For white meat it was simulated as zero. For total fruit it was -.2; this minus sign means that it's a protective factor. And for total energy, it's again a risk factor with the effect of .2 for increasing energy intake from 1,500 to 3,000 kcal.

When you use the naïve model, with means of the 24 hour recalls, all estimated log odds ratios changed. For red meat, for total fruit, and for energy, the estimated log odds ratios are attenuated. For total fruit, it's not that much; for red meat and for total energy it's about one third of what it should be, so it's quite a bit of attenuation.

Now, this is the interesting result for white meat. Remember, the true effect was zero. If you use the naïve model, you estimate the true effect as a small effect, but it's statistically significant because it exceeds this standard error by a big factor. This could take place in multiple risk models with multiple dietary components. It's the case when you create something, some effect, from nothing.

Now, what happens if we use the enhanced regression calibration predictor? In all cases, we get the results which are very, very close to true results. For example, for red meat it's .395 instead of .4 and the differences in no case are statistically significantly different from zero. So, in other words, we almost precisely recover the true log odds ratios.

SUMMARY & DISCUSSION

Slide 56

Summary ….

Summary (1)

- Developed methodology addresses major challenges for multivariate modeling of short-term unbiased measurements of dietary intakes by allowing
  - Excess zeros in episodically-consumed dietary components
  - Skewed distributions of positive intakes
  - Correlations among positive intakes of different dietary components
  - Correlations of facts of consumption of episodic components among themselves and with consumption amounts of other dietary components

Slide 57

Developed methodology addresses major challenges for multivariate modeling of short-term unbiased measurements of dietary intakes by allowing excess zeros in episodically consumed dietary components, allowing skewed distributions of positive intakes for episodically consumed dietary components or just all intakes for daily consumed components, allowing correlations among positive intakes of different dietary components, and allowing correlations of facts of consumption of episodic components among themselves and with consumption amounts of other dietary components.

Summary (2)

- New measurement error model is highly non-linear with multiple correlated latent variables and structured covariance matrix
- The model is fitted using Markov Chain Monte Carlo technique implemented in SAS
- Developed methodology allows for rigorous regression calibration correction for measurement error when repeat short-term dietary assessment methods are used as the main instrument in the study, alone or in combination with FFQ
- New methodology allows for rather flexible risk models with covariates on transformed scales

Slide 58

The new measurement error model that we discussed is highly nonlinear with multiple correlated latent variables and a structured variance-covariance matrix. The model is fitted using the Markov Chain Monte Carlo technique, which was implemented in SAS. And the developed methodology allows for rigorous regression calibration correction for nondifferential measurement error when repeat short-term dietary assessment methods are used as the main dietary assessment instrument in the study, alone or in combination with FFQ. And the new methodology allows for rather flexible risk models, which include transformation of covariates.

Discussion (1)

- We considered episodically-consumed dietary components that are eventually consumed in the long run
- What about never consumers?
  - Model could be extended to include never consumers for multiple dietary components
  - The extension is currently under development

Slide 59

Now, we considered episodically consumed dietary components which are not consumed by everyone almost every day, but which are eventually consumed in the long run. What about never consumers? There are some dietary components which are surely never consumed by some subjects. An example would be alcohol intake. So, actually, the model that we considered could be relatively easily extended to include never consumers and this extension is currently under development. The only difficulty is in fitting this model, but with Markov Chain Monte Carlo, it could be done. We are working on it as we speak.

Discussion (2)

- Developed methodology is based on the important assumption that a repeat short-term instrument is **unbiased** for true usual dietary intake
- 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 of other dietary components

Slide 60

And the most important thing is that the developed methodology is based on the assumption that repeat short-term measurements are unbiased for true usual dietary intake. And in my considered application, such unbiased measurements were done by 24 hour recall.

On the other hand, studies with recovery biomarkers that use doubly labeled water for energy intake, urinary nitrogen for protein intake, and urinary potassium for potassium intake have demonstrated some biases in 24 hour recalls, admittedly not as large biases as, for example, in FFQ but biases nevertheless, suggesting that there could be possible biases in the reporting of other dietary components.

Discussion (3)

- Our preliminary simulations based on OPEN biomarker study suggests that, in spite of biases, using repeat 24HRs in the developed methodology on average leads to better results than no correction for measurement error

- Using more precise short-term instruments, such as automated 24HR, in nutritional epidemiology is therefore a step forward toward better understanding of diet-health outcome relationships

Slide 61

We did some preliminary calculations or simulations based on the OPEN biomarker study, and those calculations suggest that in spite of those biases using repeat 24 hour recalls with the developed methodology, on average at least, leads to better results than no correction for measurement error at all, which means that using those more precise short-term instruments such as automated 24 hour recalls in the analysis of cohort studies in nutritional epidemiology has the potential of being a step forward toward better understanding of diet-health outcome relationships.

QUESTIONS & ANSWERS

Moderator: Sharon Kirkpatrick

Please submit questions using the *Chat* function

Slide 62

Thank you, Victor. We'll now move on to the question and answer period.

Measurement Error Webinar 12 Q&A

Question: **There is a question to clarify the tables from the simulation study. So if we could go back to slide 55, could you clarify what is meant by the mean 24 hour recall approach?**

So what we did was we took 24 hour recalls, or reported data from two 24 hour recalls, and we added them up and divided by two. So if a person would report two zeros, this mean would be zero. If some amount would be reported on day one, for example, and zero on day two, it would be half of the reported amount on day one, and so forth. So this is the so-called naïve approach—no adjustment for measurement error. You take an average amount—with two 24 hour recalls it's the average of the two reported days. And you use it as covariates in the risk model. *(V. Kipnis)*

Early on in your talk you mentioned that regression calibration doesn't fully restore the power that's lost due to measurement error. So to what extent does it restore power?

That's an interesting question. You may have heard in previous webinars that regression calibration basically uses the same information as the naïve model; in other words, it uses observed or measured dietary intake and the vector of covariates, Z. And because of that, the power would not be increased. It would be true in many cases; it would theoretically be true in a case of the univariate dietary exposure when the regression calibration is linear. It would not be the case when the regression calibration is nonlinear or when the risk model contains several dietary components, the reason being that when we talk about regression calibration, we use the conditional mean of a certain dietary component given all other dietary components. So information overall is the same, but for each dietary component it's packed in a different fashion. So in those multivariate models with many error-prone covariates, when the regression calibration is nonlinear there could be some increase.

Now, with enhanced regression calibration, there could be additional increase in power because now you bring in additional information, components of vector C. How can one judge how much increase would be there? Well, it depends on the precision of the regression calibration predictor. When Doug Midthune in webinar 10 gave his lecture, he measured the precision with R squares; remember, this is a squared correlation between the regression calibration predictor and true intake. So if those R squares for all involved dietary components are relatively large, close to 1 ideally, it would mean that we may restore a significant

amount of lost power. If they are not, it would mean that the power remains the same or close to being the same. *(V. Kipnis)*

Did you also compare the model that you presented today to a model without correlations, and did this make any big difference?

The model without correlation—there are many correlations involved. Remember, there are correlations among person-specific random effects, among u's. There are correlations among ε's on each given day. So the first correlations—they are showing the relationship between usual intakes. The second type of correlations shows correlations among intakes or fact of having intakes on a particular day. So did we use the comparison when we wouldn't allow such correlations to simplify the model? Not yet. I should mention that this is cutting-edge research, at least in our group. And so what I presented is hot from the oven. We haven't done everything that we wanted to. We're in the process of doing it. But it's actually a good question. If not allowing those correlations which would simplify the model produces similar results, we would be thrilled to use a simplified model. My gut feeling is, though, that it won't be the case.
(V. Kipnis)

Continuing with the note of this being a new model, can you discuss your plans for further developing or evaluating the method? And you did mention that there currently aren't any cohort studies that you could use, but could you test the method using data from NHANES or some other data source like that?

Let me start with NHANES first. In our *Biometrics* paper we did consider an example from NHANES and it was a univariate model. By univariate, I mean it was only one dietary exposure of interest measured with error. It was fish intake and outcome was mercury level. NHANES is a cross-sectional study. Any cross-sectional study has many problems with trying to estimate diet and health outcome relationships. So in principle, one can try and do it in NHANES but I am rather reluctant to consider cross-sectional studies for several reasons, one being that measurement error may not be nondifferential because of some outcomes that people may know about, like blood pressure. For example, when some people know they have elevated blood pressure, they may change their dietary intake because probably they would know about their blood pressure before the administration of 24 hour recalls. So whether their 24 hour recall therefore relates true usual intake before their blood pressure got elevated or not is a big question.

So now about our plans: There wasn't such a question yet, but I don't want you to think that our methodology is the only methodology to handle unbiased measurements of dietary intakes. Actually, in a simple situation, one can use a BLUP. Now, imagine that I want to do it. In theory, the BLUP could be used when the measurement error is classical. What people usually do is they transform the data to a scale where the error is more or less classical and then they disregard excess zeros in episodic components or the fact that on the transformed scale the instrument is not unbiased anymore because it was assumed to be unbiased on the original scale. So there would be lots of approximation involved, but the model would be simple. You don't have to do numeric integration.

Everything exists in closed form. So one thing that we would like to do is to compare this simplified approach with our methodology just to show that this methodology was worth developing. I did present the results of such a comparison in my webinar 8 when 24 hour recall was used as a reference instrument. That showed that linear regression calibration, which would be an equivalent of this BLUP approach, sometimes fails to correctly adjust for measurement error. Whether the same would take place here—I hope so but we're going to try it.

And then there was a suggestion about trying a more simplified model without correlations of either type. We're going to try this as well. *(V. Kipnis)*

You mentioned in the last few slides looking at OPEN that those results suggest that this approach is better than a naïve approach, so there's a question about the within-individual variation in biomarkers like doubly labeled water and urinary protein or the indicator that we use for protein, urinary nitrogen. So given that, what kind of study would you suggest to validate or test the method? Do you have any concerns about that within-individual variation?

Strictly speaking, individual variation matters only when you estimate the regression calibration predictor. It matters because the precision of the regression calibration predictor would depend on this individual variation when you estimate the parameters of the measurement error model. It could be an important consideration when you use short-term instruments as reference instruments because if you use them in a calibration substudy, calibration substudies are usually smallish and so your estimation of the parameters of the measurement error model may not be very precise. It would jeopardize the whole exercise.

In this particular case that I considered today, the short-term instruments are applied as the main instrument to everyone. And so even given the relatively high variation in, say, urinary nitrogen and urinary potassium, it should not jeopardize the estimation of measurement error parameters in the model. And so I think that in this case we're more or less safe. The most important thing is that the recovery markers that are unbiased for true usual intake have been established in many feeding studies as well as from physiologic research. And those are the distinct characteristics. The unfortunate thing is that we have only a few such recovery biomarkers. But there is some hope. Lawrence Freedman in webinar 11, I believe, presented some results with other types of biomarkers, which have stable relationships with dietary intake. They could be called predictive biomarkers. So they may be helpful in the same process of validation. We want to know more about them. So it's very important from my point of view to do feeding studies. I know they are expensive, but they are necessary I think. I know of one such study which is going to be done by the researchers of the Women's Health Initiative. *(V. Kipnis)*

There are a couple of questions about whether you can apply this method if you only have one recall, so, for example, if you had one recall and an FFQ.

Well, that's another thing which I didn't mention, which we are going to look at. Strictly speaking, I don't know of any study where a short-term instrument would be the main instrument applied only once. But if somebody would like to save some money and do it, the problem with this is that now we don't have longitudinal data. So we won't be able to distinguish between person-specific random effect and within-person variation. The model would be simplified tremendously. It would now only contain fixed effects. On the other hand, because within-person variation, which I called ε, now contains as part of it this person-specific random effect, which couldn't be separated from it, it means that we would have to allow all ε's to be correlated. So on the one hand, the model is simplified; on the other hand it gets a little bit more complex. Whether it could be fit; whether it's identifiable; whether it will, if identifiable, produce meaningful results, I don't know at this point in time. But I would like to look into it. *(V. Kipnis)*

Could you speak to what type of measurement errors you're correcting for with this new calibration approach—both random and systematic?

Remember, the main assumption is that the short-term instrument that we use as the main instrument of dietary assessment in this study is

unbiased, which means that the systematic error doesn't exist. It's random variation. In enhanced regression calibration we use FFQs, which as we well know now, contain both types of measurement error, random and systematic. So we use an unbiased instrument and FFQ data to do enhanced regression calibration, basically to adjust for random error in short-term instruments using the FFQ as an additional source of information. *(V. Kipnis)*

Does using the Markov Chain Monte Carlo or MCMC mean that you propose a Bayesian approach?

Well, as some of you may know, Markov Chain Monte Carlo or MCMC is usually used within the Bayesian paradigm but it doesn't have to be. So the answer to this question is no. We use a frequentist interpretation and it's based on the well-known asymptotic equivalence of the maximum likelihood and MCMC, so it's a pseudo-Bayesian paradigm. Could it be used in a full Bayesian approach? Yes, it could; we just haven't done it. *(V. Kipnis)*

So thinking back to webinar 10 when Doug Midthune discussed designing studies and he mentioned that an optimal or close to optimal design for the case of one component would include four 24 hour recalls plus an FFQ, does that also apply to the multivariate case?

First of all, what Doug has presented was for absolute intakes as well as energy-adjusted intakes, or residuals. So in this sense, it's sort of like a bivariate model. But would it hold for the multivariate model? It's one of the items on the to do list, but my sort of intuitive understanding is that unless after energy adjustment, say using the residual or density method, the energy-adjusted intakes are highly correlated—unless this happens the results of multivariate modeling would be pretty close to what Doug has presented. So I would expect that in most cases four 24 hour recalls and an FFQ would be close to the optimal design, but we need to verify it. *(V. Kipnis)*

[This page intentionally blank.]

measurement ERROR webinar series

This concludes our Webinar series.

Thank you for participating.

For access to series archives and supporting materials, please visit:
riskfactor.cancer.gov/measurementerror/sessions

Slide 63

Thank you, Victor, and thank you to all of the presenters and collaborators who have contributed to this series. Also, thank you to our audience for joining us over the past several weeks. We have appreciated the questions submitted to our Q&A sessions and the comments sent via email and we look forward to continuing the dialogue at upcoming meetings, including the International Conference on Diet and Activity Methods in Rome next year. Please note that you can find recordings of all of the presentations on the Web site mentioned earlier. We are in the process of posting the slides, with notes, along with the recordings on our main Web site. And with that, we will end today's session and close out the series. Goodbye.