Nutrition Epidemiology
Implications of Error for Various Applications

Laura Caulfield, PhD
Professor
Center for Human Nutrition
Implications of errors in our variables

• day to day variation in dietary intake is real but when trying to describe differences across people, it is a source of error

• The same is true when thinking about measurement error (analytic error) and other error in biochemical measures

• Validity coefficients for FFQ also indicate significant error

• The amount of error influences many aspects of study design/protocol

• The nature and degree of error have implications for our measures of association and the inferences we wish to draw
Implications of error

A. Estimating the usual nutrient intake of a person
   • Within-individual variation acts as random error on estimates of the usual intake of individuals with respect to a nutrient, making it less precise than expected
   • Using the formula:

\[ N = (Z \times CV_w / D)^2 \]

Where

- \( N \) = number of days of intake/person needed (replicates)
- \( Z \) = normal deviate for the % time the measured intake should be within a specified limit (e.g., alpha = .05; 95% time, \( Z = 1.96 \))
- \( D \) = specified limit (e.g. within +- 20%)
- \( CV_w \) = within-individual coefficient of variation
  (standard deviation \( w \) / mean) * 100
Examples

Given CV$_w$ for energy and cholesterol are 20% and 50%, respectively, and taking values of $Z = 1.96$ and $D = 20\%$, we find that:

For Energy:  
\[ N = \left( \frac{1.96 \times 20\%}{20\%} \right)^2 \]
\[ = 3.8 \text{ days/person} \]

For Cholesterol:  
\[ N = \left( \frac{1.96 \times 50\%}{20\%} \right)^2 \]
\[ = 24 \text{ days/person} \]

• Therefore, for example, we would need 24 days of intake/person in order that 95% of the time our estimates of the person's cholesterol intake would fall within (plus/minus) 20% of the true usual value.

• Alternatively, one can calculate that given the within-individual variation in cholesterol intake, if you had one day's estimated intake, 95% of the time your estimate if the person's usual cholesterol intake would be plus or minus 98% of the true value (For energy, you would be within 39%).
Number of repeated days needed per person for 95% of observed values to lie within specified % of true mean

<table>
<thead>
<tr>
<th>Nutrient</th>
<th>Number of days needed to lie within specified % of true mean</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Within person coefficient of variation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total fat</td>
<td>38.4</td>
<td>57</td>
<td>14</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Calorie adjusted</td>
<td>19.8</td>
<td>15</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>62.2</td>
<td>149</td>
<td>37</td>
<td>17</td>
<td>9</td>
</tr>
<tr>
<td>Calorie adjusted</td>
<td>61.5</td>
<td>145</td>
<td>36</td>
<td>16</td>
<td>9</td>
</tr>
<tr>
<td>Sucrose</td>
<td>60.3</td>
<td>140</td>
<td>35</td>
<td>16</td>
<td>9</td>
</tr>
<tr>
<td>Calorie adjusted</td>
<td>50.1</td>
<td>96</td>
<td>24</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>Vitamin A</td>
<td>105.0</td>
<td>424</td>
<td>106</td>
<td>47</td>
<td>26</td>
</tr>
<tr>
<td>Calorie adjusted</td>
<td>104.7</td>
<td>424</td>
<td>106</td>
<td>47</td>
<td>26</td>
</tr>
</tbody>
</table>

Adjusted for total energy using regression analysis
Implications of error

• Using this formula, one can calculate how many days (replicates) per person are needed to obtain an estimate of an individual’s nutrient intake with a given level of precision.

• One can also calculate the level of precision that one has of an individual’s estimated nutrient intake, given specified sampling conditions.

• One can also apply this formula to consider non-dietary measures, e.g., biochemical indicator of nutrient status to obtain information at the level of the individual.

• This need for replicates for an individual mimics the situation you have in cohort studies, e.g., where you are interested in having a precise individual estimate of the exposure but you are collecting this information on N individuals (more to come).
Implications of error

B. Estimating the usual nutrient intake of a group/population

• Within-individual variation acts as random error on estimates of the usual intake of individuals with respect to a nutrient, making the estimate of the distribution less precise than expected.

• In this situation, we can increase the precision of the sample estimate by either increasing the sample size or increasing the number of replicates per person.

• Using the formula:

\[ D = Z \times \sqrt{\frac{(CV_b^2)}{N} + \frac{(CV_w^2)}{N \times n}} \]

Where

- \( N \) = number of individuals (sample size)
- \( n \) = number of days of intake/person needed (replicates)
- \( Z \) = normal deviate for the % time the measured intake should be within a specified limit (e.g., alpha = .05; 95% time, \( Z = 1.96 \))
- \( D \) = specified limit (e.g. within +- 20%)
- \( CV_w \) = within-individual coefficient of variation (standard deviation \( w \) / mean) * 100
- \( CV_b \) = between-individual coefficient of variation (standard deviation \( b \) / mean) * 100
Examples

Given a $C V_w$ of 23% and a $C V_b$ of 25% for energy (ratio ~ 0.85), we could obtain an estimate of the population's usual energy intake to within plus or minus 10% in the following ways:

<table>
<thead>
<tr>
<th>N</th>
<th>n</th>
<th>Size of “analytic” sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>39</td>
<td>1</td>
<td>39</td>
</tr>
<tr>
<td>32</td>
<td>2</td>
<td>64</td>
</tr>
<tr>
<td>29</td>
<td>3</td>
<td>72</td>
</tr>
<tr>
<td>28</td>
<td>4</td>
<td>112</td>
</tr>
</tbody>
</table>
Examples

- Compare the previous results with a $CV_w$ of 112% and a $CV_b$ of 50% for retinol (ratio ~ 5), we could obtain an estimate of the population's usual energy intake to within plus or minus 10% in the following ways:

<table>
<thead>
<tr>
<th>N</th>
<th>n</th>
<th>Size of “analytic” sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>502</td>
<td>1</td>
<td>502</td>
</tr>
<tr>
<td>261</td>
<td>2</td>
<td>522</td>
</tr>
<tr>
<td>181</td>
<td>3</td>
<td>543</td>
</tr>
<tr>
<td>141</td>
<td>4</td>
<td>564</td>
</tr>
</tbody>
</table>
Implications for Sample Size, Sampling

• Here, within individual variation is an additional source of variation, the observed variance is inflated.
• The degree of inflation is positively associated with amount of error variance.
• You may be able to address this directly and obtain “true between subject- variance” through sampling (N number of people, n number of replicates).
• This may be important if you have limited ability to increase N (clinical patients, e.g.), but can do replicates on those you find.
• One can consider the costs of increasing N with the costs of increasing n.

These points are made using diet as an example but are generalizable to other exposures with random error.
Estimating the usual distribution

• We want to characterize the distribution of something:
  – What is the usual protein intake of the US adult population?
  – What is the distribution of plasma retinol among children in Peru?
• We want to Monitor a distribution over time
  – Is it changing, are we making progress?
• We want to compare distributions
  – Is the distribution different in Population A from Population B?
• We want to calculate the prevalence of a condition by applying a cut-point to the distribution
• In all of these instances, we want the true variance and because of random error the variance is inflated undermining our ability to answer the questions we have
Impact of the distribution on prevalence

- With increasing replicates the SD you observe is the true SD
- If you have data on between and within CV or variances, you can use that information to shrink the distribution
- You may have this from an ANOVA in which you have evaluated sources of error and have $S_b^2$
- Surveys such as NHANES collect replicate(s) on everyone or on a subsample for this purpose, depending on the source of error

**FIG 2.** Simulation of impact of random error on observed intake distributions. The assumptions are that both between- and within-person variations have a CV of 25% and that both are normally distributed. Also shown is the impact that days of data collected would have on the proportions of individuals with low (< 50 units) or high (> 150 units) intakes.
Illustration of impact of variance inflation due to random error on prevalence (bias is dependent on magnitude of random error and the $P_{true}$).

$P_{obs} > P_{true}$

Source: Biehl et al., 2013
Adjusting the distribution of the data

• You have estimated the error by calculating the Reliability of the indicator

\[ R = \left[1 - \frac{s^2_{\text{unreliability}}}{s^2_{\text{observed}}}\right] \times 100 \]

e.g. = 90%

• Using this information you can “correct” or adjust the distribution of the data thereby allowing you to get an unbiased estimate of the prevalence

• This is done, e.g., by calculating the adjusted values of the indicator in the sample
Equation for adjustment

\[ \text{adjustedvalue}_i = \bar{x}_{true} + [\text{value}_i - \bar{x}_{true}] \times \left[ 1 - \left( \frac{s^2_{unreliability}}{s^2_{observed}} \right) \right] \]

- For example, a value that falls at -1 unit below \( x_{true} \) will have an adjusted value equal to -0.90, and a value -2 units below the \( x_{true} \) will have an adjusted value equal to -1.8.

- Each value is having 10% of the distance it is from the mean “shaved off of it” or it is being shrunk.

- The end result of this is a distribution of indicator X with \( X_{adjusted} = X_{true} \) (i.e. no systematic error present) AND \( SD_{adjusted} = 0.90 \times SD_{observed} \)

- Now, using the characteristics of the adjusted distribution of indicator X, an unbiased estimate of the prevalence of the condition can be calculated.
Correcting distributions for error

• If we collect information on magnitude and sources of error, we can use that information to make post hoc adjustments to our data

• If we want to estimate the mean and variance of a distribution of values for an indicator assessed with error, we can use estimates of R or Validity coefficients to “correct” our distribution for error

  \[ SD_{true} = \sqrt{Var(TM)} \] from ANOVA

  \[ SD_{true} = \sqrt{Var(OM) - (Var(ME) + Var(OE))} \]

  \[ SD_{true} = R \times SD_{observed} \]

  \[ SD_{true} = "Validity coefficient" \times SD_{observed} \]
A variety of methods have been developed to estimate the population usual nutrient intake and true variance.

- CSIDE (Iowa State Univ: see Nusser, or Dodd)
- NRC method
- NCI method and improved NCI method (Tooze 2006; 2010)
- Jahns et al (2005) using extant or reported values
- Yanetz et al, to remove bias on mean using biomarkers
- Others

- These methods have also been extended to estimate usual food intake distributions by combining information from FFQ and 24 hour recall data.
Estimation of usual intake distributions: General procedures for advanced methods

• Using the 24-hour recall data you can estimate the distribution of nutrient intake adjusting for within-subject variability if you have this information on at least some individuals in the sample:
  – You transform the distribution to approximate normal
  – You shrink the distribution using information on within-between subject variance to create an idealized distribution
  – You can model intake as a function of additional specific covariates: age, gender, ethnicity,
  – You back transform to original scale

• For most nutrients, the Pr(intake/consumption) is 1, so this is it

• When estimating food intakes and nutrients for which Pr(consumption) < 1 (in prior section) a 2 step procedure is needed where you first combine frequency data from FFQ with quantities from 24 hour recall and then the above is step 2.
Approaches for adjustment

\[ \text{adjusted value}_i = \frac{x_{\text{true}}}{\text{mean}_{\text{observed}}} + \left[ \text{value}_i - x_{\text{true}} \right] \times \left[ 1 - \left( \frac{s^2_{\text{unreliability}}}{s^2_{\text{observed}}} \right) \right] \]

- Here above and on earlier slides, we used Reliability as a shrinkage factor
- Another formulation proposed is to use the square root of ratio of the variances between the biomarker (true variance) and intake (observed variance) as the shrinkage factor. For example, taking the square root of the variance of energy intake from deuterium divided by the variance of observed energy intake (constrained by N for deuterium)

- Suppose there is bias and observed mean is not true mean, Yanetz et al (2008) recommend estimating the ratio of the true mean (mean according to reference method or biomarker) to the observed mean and using this to shift the mean. Estimate within your study or use extant information
- Adjusted mean = \( \left( \frac{\text{mean}_{\text{true}}}{\text{mean}_{\text{observed}}} \right) \times \text{mean}_{\text{observed}} \)
Effect of methods on energy intake distributions (Yanetz et al., 2011)

Traditional (observed): 1595 (666)
NRC (adjusting for within-subject variance): 1555 (428)
NRC-B (biomarker) with bias adjustment: 1714 (382)
Biomarker suggests higher protein intake than FFQ. None of the 5 methods of calibrating the protein through adjustment of energy intake was “the best” method. Goldberg method identified with greatest accuracy (in PABA subset) but this method involved retaining 44/162 subjects!
Implications of error

C. Measures of association
   – Correlation coefficients
   – Regression coefficients
   – Odds ratios (OR) and relative risks (RR)
Effects on measures of association between dietary intake (X) and outcomes of interest (Y)

Pearson Correlation Coefficient

\[
r_{\text{observed}} = \frac{r_{\text{true}}}{\sqrt{\left(1 + \frac{s_{WX}^2}{s_{BX}^2 n_X}ight) \left(1 + \frac{s_{WY}^2}{s_{BY}^2 n_Y}\right)}}
\]

Where

\( r_{\text{true}} \) the true correlation between X and Y
\( r_{\text{observed}} \) the observed correlation between X and Y
\( s_{WX}^2 \) the within-person variance for X
\( s_{BX}^2 \) the between-person variance for X
\( n_X \) the number of replicates/person
Using Intra-class correlation (ICC) to do the same thing

The ICC assesses the reproducibility of repeat measures. Information on the ICC of each variable (x and y) can be used to calculate the ratio $s_w^2/s_b^2$ because:

$$\frac{s_w^2}{s_b^2} = \frac{(1-\text{ICC})}{\text{ICC}}$$

Either method can be extended for Spearman correlation coefficients as well.
Implications of error

Correlation coefficients (r)

• The correlation coefficient that you observe is biased toward null due to the presence of random error
• The only way to diminish the bias is to increase replicates or reduce the ratio by some other means
• Increasing sample size will allow you to be able to detect the biased coefficient as statistically significant
• De-attenuating r means you have to re-calculate the variance of r, and the confidence interval will be wider than before because you are only estimating error
Linear regression of Protein Intake

Protein Intake (g/day) = \( a + b \cdot \text{Protein} \)

Y is continuous and normally distributed, or transformable to normal

B = change in Y per unit change in X (here, grams of protein/day), is the exposure outcome relation you are interested in
Implications of error

Regression (OLS)

- The beta coefficient that you observe is biased toward null due to the presence of error

- If $b_{\text{true}} = .50$ and ratio = 4 and you have one replicate, $b_{\text{observed}} = 0.10$

- The only way to diminish the bias is to increase replicates or reduce the ratio by some other means
Suppose there is systematic within subject error? Regression calibration

• If you have results on the relationship between the exposure and a reference/gold standard you can use this information in a linear regression model to estimate the degree of bias

\[ X_{\text{true}} = a_0 + a_1 X_{\text{observed}} \]

• In this model, you are interested in \( a_1 \) which is the change in true \( X \) given a one unit change in observed \( X \)

• The \( a_1 \) estimated above can be used to calibrate the beta relating exposure to outcome

• \( \text{Beta}_{\text{true}} = [\text{Beta}_{\text{obs}}/a_1] \)
Implications of error

• Odds Ratios (OR) and Relative Risks (RR)
  – The impact of random measurement error and other error is to bias toward the null the estimated OR or RR, because the error results in misclassification of exposure status
• There are extensions to other risk estimates (HR)
Example: Let the true relation between exposure and disease defined by a 2 X 2 table be as follows:

<table>
<thead>
<tr>
<th></th>
<th>Exposed</th>
<th>Unexposed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diseased (D)</td>
<td>A = 400</td>
<td>B = 600</td>
<td>1000</td>
</tr>
<tr>
<td>Well (W)</td>
<td>C = 200</td>
<td>D = 800</td>
<td>1000</td>
</tr>
</tbody>
</table>

The Odds Ratio (OR) = \((ad) / (bc) = (400*800) / (200*600) = 2.67\)
### Misclassification in a 2 X 2 table

<table>
<thead>
<tr>
<th></th>
<th>Exposed</th>
<th>Unexposed</th>
<th>Exposed</th>
<th>Unexposed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diseased</td>
<td>A = 240 + 60</td>
<td>B = 540 + 160</td>
<td>Diseased</td>
<td>A = 300</td>
<td>B = 700</td>
</tr>
<tr>
<td>Well</td>
<td>C = 120 + 80</td>
<td>D = 720 + 80</td>
<td>Well</td>
<td>C = 200</td>
<td>D = 800</td>
</tr>
</tbody>
</table>

The Odds Ratio (OR) = \((ad) / (bc) = (300 \times 800) / (200 \times 700) = 1.71\)
Implications of error

- In a logistic model, we estimate a beta coefficient and then exponentiate the beta to obtain the risk ratio (OR or RR)
- The beta we estimate is biased toward null because the associated X variable contains random error and so is the risk ratio estimated from it (biased)
- If we can estimate the relation between the observed values of X and the true values of X, then we can estimate the degree of bias
Again, regression calibration

We can estimate the relation between the observed values of $X$ and the true values of $X$, then we can estimate the degree of bias.

We can do that through a validation study, or calibration sub-study. On a sample of subjects, get data on $X_{\text{true}}$ and $X_{\text{observed}}$

$$X_{\text{true}} = a_0 + a_1 X_{\text{observed}}$$

where $X$ is, e.g., nutrient intake.
Regression calibration model

We know that

\[ \beta_{\text{true}} = \frac{\beta_{\text{observed}}}{a_1} \]

Therefore the regression coefficient \( a_1 \) provides us with the means to estimate the degree of bias towards the null in RR or OR, this is sometimes called the attenuation factor

\[ \text{RR}_{\text{corr}} = e^{\beta_{\text{observed}}/a_1} \]
\[ \text{RR}_{\text{observed}} = (\text{RR}_{\text{true}})^{a_1} \]
Nitrates and nitrites from food: results from AARP validation study (Inoue-Choi et al., 2015)

<table>
<thead>
<tr>
<th></th>
<th>Correlation coefficient</th>
<th>Attenuation factor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Men</td>
<td>Women</td>
</tr>
<tr>
<td>Total nitrate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unadjusted</td>
<td>0.53</td>
<td>0.48</td>
</tr>
<tr>
<td>Adjusted</td>
<td>0.59</td>
<td>0.57</td>
</tr>
<tr>
<td>Total nitrite</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unadjusted</td>
<td>0.45</td>
<td>0.37</td>
</tr>
<tr>
<td>Adjusted</td>
<td>0.59</td>
<td>0.58</td>
</tr>
<tr>
<td>Animal nitrite</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unadjusted</td>
<td>0.61</td>
<td>0.47</td>
</tr>
<tr>
<td>Adjusted</td>
<td>0.64</td>
<td>0.52</td>
</tr>
<tr>
<td>Plant nitrite</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unadjusted</td>
<td>0.52</td>
<td>0.48</td>
</tr>
<tr>
<td>Adjusted</td>
<td>0.68</td>
<td>0.68</td>
</tr>
</tbody>
</table>

Attenuation factors are the beta coefficients from a regression of true intake (from 24HR recall) = FFQ intake (a measurement error model). Adjusted means adjusting for energy using residual method (section 7).
### Table 12–3. Observed relative risks for different levels of validity in the measurement of exposure

<table>
<thead>
<tr>
<th>$\gamma^b (\hat{\alpha}_i)$</th>
<th>True relative risks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.5</td>
</tr>
<tr>
<td>0.2</td>
<td>1.08</td>
</tr>
<tr>
<td>0.3</td>
<td>1.13</td>
</tr>
<tr>
<td>0.4</td>
<td>1.18</td>
</tr>
<tr>
<td>0.5</td>
<td>1.22</td>
</tr>
<tr>
<td>0.6</td>
<td>1.28</td>
</tr>
<tr>
<td>0.7</td>
<td>1.33</td>
</tr>
<tr>
<td>0.8</td>
<td>1.38</td>
</tr>
<tr>
<td>0.9</td>
<td>1.44</td>
</tr>
<tr>
<td>1.0</td>
<td>1.50</td>
</tr>
</tbody>
</table>

\(^a\) $RR_0 = (RR_1)^\gamma$ where $RR_0$ is the observed relative risk and $RR_1$ is the estimated true relative risk.

\(^b\) The regression coefficient for the true measure on the surrogate measure or (when both measures have the same standard deviation) the correlation coefficient between them.
# Reported versus true intakes of fat (Freedman et al, 1990)

**Table 1**  
*Distribution of reported and true intakes of fat, measured by percent calories from fat*

<table>
<thead>
<tr>
<th>Percent consuming less than</th>
<th>Reported percent calories from fat</th>
<th>True percent calories from fat</th>
</tr>
</thead>
<tbody>
<tr>
<td>25%</td>
<td>5</td>
<td>0.5</td>
</tr>
<tr>
<td>30%</td>
<td>16</td>
<td>6</td>
</tr>
<tr>
<td>35%</td>
<td>36</td>
<td>29</td>
</tr>
<tr>
<td>40%</td>
<td>62</td>
<td>68</td>
</tr>
<tr>
<td>45%</td>
<td>83</td>
<td>93</td>
</tr>
<tr>
<td>50%</td>
<td>95</td>
<td>99.4</td>
</tr>
</tbody>
</table>

* p < .05
### Attenuation of relative risk (RR) estimates

<table>
<thead>
<tr>
<th>Gradient</th>
<th>$&lt;$ 25</th>
<th>25 – 32.5</th>
<th>32.5 - 40</th>
<th>40 – 47.5</th>
<th>$&gt;$47.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>1.0</td>
<td>1.09</td>
<td>1.18</td>
<td>1.27</td>
<td>1.36</td>
</tr>
<tr>
<td>Medium</td>
<td>1.0</td>
<td>1.16</td>
<td>1.32</td>
<td>1.48</td>
<td>1.64</td>
</tr>
<tr>
<td>Large</td>
<td>1.0</td>
<td>1.33</td>
<td>1.67</td>
<td>2.00</td>
<td>2.33</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reported percent calories from fat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gradient</td>
</tr>
<tr>
<td>Small</td>
</tr>
<tr>
<td>Medium</td>
</tr>
<tr>
<td>Large</td>
</tr>
</tbody>
</table>
Table 5

For fat and colorectal cancer, numbers required to detect a significant effect at 5% level with 90% power using trend test over 5% calories from fat groups: <25%, 25–32.5%, 32.5–40.0%, 40.0–47.5%, and ≥47.5%.

<table>
<thead>
<tr>
<th>Relative risk*</th>
<th>Correlation between measured and true exposure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.60</td>
</tr>
<tr>
<td>Small</td>
<td></td>
</tr>
<tr>
<td>Cases‡</td>
<td>14,500</td>
</tr>
<tr>
<td>Cohort§</td>
<td>1,455,000</td>
</tr>
<tr>
<td>Moderate</td>
<td></td>
</tr>
<tr>
<td>Cases</td>
<td>5,600</td>
</tr>
<tr>
<td>Cohort</td>
<td>562,000</td>
</tr>
<tr>
<td>Large</td>
<td></td>
</tr>
<tr>
<td>Cases</td>
<td>2,000</td>
</tr>
<tr>
<td>Cohort</td>
<td>201,000</td>
</tr>
</tbody>
</table>

* Over 5 groups (<25, 25–32.5, 32.5–40.0, 40.0–47.5, and ≥47.5). Small: 1.00, 1.09, 1.18, 1.27, and 1.36; moderate: 1.00, 1.16, 1.32, 1.48, and 1.64; large: 1.00, 1.33, 1.67, 2.00, and 2.33.
† Correlation = 1.00 represents no measurement error.
‡ Total number of cases in cohort. Calculated using Breslow and Day (15). Rounded to the nearest 100.
§ Assuming 5-year follow-up; 199.3 cases per 100,000 subjects per year. Rounded to the nearest 1,000.
### Table 6

*Preventable proportion of colorectal cancer were the population to shift their fat intake to the next lowest group*

<table>
<thead>
<tr>
<th>Percent calories from fat</th>
<th>Proportion with true intake in the range</th>
<th>Relative risk gradient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Small</td>
</tr>
<tr>
<td>&lt;25.0</td>
<td>0.01</td>
<td>1.00</td>
</tr>
<tr>
<td>25.0–32.5</td>
<td>0.14</td>
<td>1.09</td>
</tr>
<tr>
<td>32.5–40.0</td>
<td>0.53</td>
<td>1.18</td>
</tr>
<tr>
<td>40.0–47.5</td>
<td>0.30</td>
<td>1.27</td>
</tr>
<tr>
<td>&gt;47.5</td>
<td>0.02</td>
<td>1.36</td>
</tr>
</tbody>
</table>

Preventable proportion (%)

Reduction in number of colorectal cancers per annum in the United States (current incidence, 147,000 per annum)

<table>
<thead>
<tr>
<th></th>
<th>11,000</th>
<th>17,200</th>
<th>27,800</th>
</tr>
</thead>
</table>

\[ \text{Preventable proportion} = \sum_{i=2}^{5} \frac{p_i (r_i - r_{i-1})}{\sum_{i=1}^{5} p_i r_i} \]

If \( p_i \) is the proportion in the \( i \)th group (\( i = 1-5 \) in ascending order of fat intake) and \( r_i \) is the relative risk.
Recommendations for study design given high amounts of measurement error

1. Clearly conceptualize the theoretical exposure, i.e., how do you think intake and disease are related, and when is the relevant period (time frame) of exposure influencing the disease process?

2. Choose your methodology accordingly

3. Conduct a validation/calibration study of the chosen methodology
   a. Assess how well the methodology measures true intake
   b. Conduct repeat measures of the methodology to assess reliability (precision)
   c. Carry this out as a sub study of the current study, that is from the same population (do it every study)

4. When analyzing the data don’t just look at the correlation coefficients – do regression analyses, ANOVA, covariance structure analyses (are errors correlated?), consider random error and bias
Calibrations studies (culled from Rosner and others)

- Consider that having 1 replicate of principal exposure on everyone may be really important/valuable (not sub-study)

- For calibration sub-studies, obtain both a reference dietary method for comparison as well as 1 or more biomarkers, have repeated recalls/records over the same exposure period as the FFQ

- Obtain replicates on all 3 (FFQ, record/recall, biomarker). ALSO measure other covariates you think will be in your “final model”

- How many subjects to have biomarker replicate data on is a function of cost, feasibility and error variance

- Utilize method of triads to understand how the 3 relate to each other, and to obtain best estimate of calibration coefficient and its variance
6. There are a lot of advanced methods developed for dealing with measurement error (see Willett, Chapter 12, Bennett et al 2017, NCI dietary validation/calibration registry, the work of Rosner or Spiegelman or Carroll)

7. Consider how sensitive are your results to assumptions you have made concerning measurement errors, and particularly about the possibility of other covariates measured with error.

8. This stuff is complex – nutritionists and statisticians should work together