Addressing error in laboratory biomarker studies

Elizabeth Selvin, PhD, MPH
Associate Professor of Epidemiology and Medicine
Co-Director, Biomarkers and Diagnostic Testing
Translational Research Community (TRC), ICTR
“Biological marker” – a cellular, biochemical, or molecular indicator an exposure; biological, subclinical or clinical process or disease indicator.

“…a measurable and quantifiable biological parameter…” (MeSH term)

Categories of biomarker measurements
- **Biosample test** (i.e., a measurement in blood, urine, or tissue)
- **Recording obtained from a person** (e.g., blood pressure, ECG)
- **Imaging test** (e.g. echocardiogram, CT scan)

See: Porta, *Dictionary of Epidemiology*; Vasan RS. *Circulation* 2006
Types of Biosample/Lab Biomarkers

- **Biomarkers of exposure – used for monitoring**
  - Environmental or toxic exposure (e.g. lead, cadmium, cotinine)

- **Biomarkers of genetic susceptibility – used for health risk assessment**
  - Genetic variants that predispose to disease (e.g. APOE)

- **Biomarkers of disease – used for screening, diagnosis, or prognosis**
  - Typically blood, urine, or tissue measurements
  - Provide information re manifestation of a disease and often represent a surrogate for clinical or pre-clinical disease
  - May serve as surrogate endpoint in trials or other studies
General framework of sources of measurement error and misclassification

Data collection

- Observer/diagnostic error
- Instrument/method error
- Reporting/transmission error
- Recording errors
- Entry errors

Data utilization
What are the sources of error in measuring a factor?

- Error due to the **person**
- Error due to the **measurement tool**
  - Examples of measurement tools: bathroom scale; blood pressure cuff (sphygmomanometer); a diet questionnaire; a laboratory assay
- Error due to the **observer**
  - Examples of observers: participant (e.g. self-report), interviewer, abstractor
- Error in **recording the measurement**
Error components

Measured value = True value + Error

Error = Bias + Random Error

Systematic component of the error
Random component of the error
Quantifying Error: Introduction to Validity and Reliability
To assess extent of random error, perform repeated measurements on the same person/sample (i.e., *replicates*).

e.g.,
- building duplicate measurements into a study
  - Same sample measured multiple times
  - Blind duplicate samples
- repeating blood collections a few weeks apart to assess short-term repeatability of lab assays
Short-term Variability in Measures of Glycemia and Implications for the Classification of Diabetes

Elizabeth Selvin, PhD, MPH; Ciprian M. Crainiceanu, PhD; Frederick L. Brancati, MD, MHS; Josef Coresh, MD, PhD

Annals of Internal Medicine
Brief Communication: Clinical Implications of Short-Term Variability in Liver Function Test Results

Mariana Lazo, MD, ScM; Elizabeth Selvin, PhD, MPH; and Jeanne M. Clark, MD, MPH

Within-Person Variability in High-Sensitivity C-Reactive Protein

Julie K. Bower, PhD, MPH
Mariana Lazo, MD, PhD, ScM
Stephen P. Juraschek, BA
Elizabeth Selvin, PhD, MPH
“An expression of the degree to which a measurement measures what it purports to measure.”

- Porta, *Dictionary of Epidemiology, 5th Ed.*, 2008
Error: bias

- Systematic difference between the true value and the measured value
  - How close is the measured value to the true value?

- In practice:
  - How close is the measured value to a “gold standard” value?
  - How close is the measured value to a “standard” value?

OR
Scenarios in epidemiologic studies

- Optimal scenario:
  - No error. Measurement = true value

- Common scenarios:
  - Random error only, ideally minimal
  - Small bias, but cannot be estimated (direction unknown)

- Manageable scenario:
  - Bias is present, but the bias can be estimated and accounted for (magnitude and direction can be evaluated and quantified)

- Problematic scenarios:
  - Substantial bias, cannot be estimated.
  - High levels of random error (with or without bias)
# Common Measures of Validity and Reliability

<table>
<thead>
<tr>
<th>Validity</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Compare to a gold standard</strong></td>
<td><strong>Compare repeated measures</strong></td>
</tr>
<tr>
<td><strong>Calculate:</strong></td>
<td><strong>Calculate:</strong></td>
</tr>
<tr>
<td>- Spearman’s/Pearson’s correlation</td>
<td>- Spearman’s/Pearson’s correlation</td>
</tr>
<tr>
<td>- Regression</td>
<td>- Regression</td>
</tr>
<tr>
<td>- Percent agreement</td>
<td>- Percent agreement</td>
</tr>
<tr>
<td>- Percent positive agreement</td>
<td>- Percent positive agreement</td>
</tr>
<tr>
<td>- Kappa</td>
<td>- Kappa</td>
</tr>
<tr>
<td>- Sensitivity/Specificity</td>
<td>- Coefficient of variation</td>
</tr>
<tr>
<td><strong>Visual displays:</strong></td>
<td><strong>Visual displays:</strong></td>
</tr>
<tr>
<td>- Scatterplot</td>
<td>- Scatterplot</td>
</tr>
<tr>
<td>- Bland-Altman plot</td>
<td>- Bland-Altman plot</td>
</tr>
</tbody>
</table>
Comparisons measurements with continuous distributions

- Calculate the *correlation coefficient* \((r)\)
  - Pearson’s
  - Spearman’s (if the distribution is not normal; based on ranks)

- Scatterplot – useful visual display
Correlation coefficient

“Measure of association that indicates the degree to which two variables have a linear relationship.”

- Porta, *Dictionary of Epidemiology*

- **Range:** -1 to +1
  - Perfect positive +1
  - Perfect negative -1
  - No correlation 0
Pearson’s correlation: examples

http://en.wikipedia.org/wiki/Correlation
Correlation coefficient: Interpretation

- Pearson’s correlation measures how close the data are to the “line of best fit” NOT fit to the “line of agreement” (i.e., y=x, 45-degree line).

- Pearson’s correlation can be a misleading measure of agreement (and its p-value is irrelevant):
  - Depends on range the data
  - Tests the null hypothesis of no linear relationship between two variables, not whether there is agreement between two measurements.
The effects of freeze–thaw on β-trace protein and β2-microglobulin assays after long-term sample storage

Stephen P. Juraschek a,b, Josef Coresh a,b,c, Lesley A. Inker d, Gregory P. Rynders e, John H. Eckfeldt e, Elizabeth Selvin a,b,*

Comparison of Two Assays for Serum 1,5-Anhydroglucitol

Regression: Y = 0.02+0.91*X
r = 0.94

Regression: Y = -0.06+1.03*X
r = 0.99

Selvin E, Rynders G, Steffes M
Measurement of HbA1c from stored whole blood samples in the Atherosclerosis Risk in Communities study

Elizabeth SELVIN,¹,² Josef CORESH,¹,²,³ Hong ZHU,³ Aaron FOLSOM⁴ and Michael W. STEFFES⁵

Glycated Hemoglobin, Diabetes, and Cardiovascular Risk in Nondiabetic Adults

Elizabeth Selvin, Ph.D., M.P.H., Michael W. Steffes, M.D., Ph.D., Hong Zhu, B.S., Kunihiro Matsushita, M.D., Ph.D., Lynne Wagenknecht, Dr.P.H., James Pankow, Ph.D., M.P.H., Josef Coresh, M.D., Ph.D., and Frederick L. Brancati, M.D., M.H.S.
Bland-Altman Plot: Method for Comparing two Measurements
Bland Altman Plot

- Compare 2 measurements graphically
- Bias
- Random error
- Same plot!
Bland Altman Plot

- How well do two measurements agree?
- Plot the difference against the mean
  - Investigate relationship between error and estimate of the “true value”
- Add a zero line
  - Is the mean of the difference different from zero?
- Add “limits of agreement” =
  - mean difference +/- 2*SD_{difference}
Bias in NHANES serum creatinine

- Calibration of serum creatinine values to standardized creatinine and commutability of serum creatinine across surveys are essential for correctly estimating kidney function and kidney disease in the population.

- Systematic (upwards) bias was present in serum creatinine measurements in NHANES.

- We directly re-calibrated serum creatinine in NHANES to an assay traceable to gold standard methods:
  - Random sample of 200 specimens from each survey
  - Analyzed for serum creatinine with assay traceable to gold standard reference methods
  - Compared ‘old’ NHANES method to new ‘gold standard’ method
Bias in NHANES serum creatinine

"gold standard" (CCRL)

Pearson’s correlation \((r) = 0.95\)

“serum creatinine measured in the NHANES survey”
Bias in NHANES serum creatinine

"Gold standard" – original NHANES

Mean (= ("gold standard" + original) / 2)

Substantial bias

-0.231 mg/dL

Elizabeth Selvin, PhD, MPH,1,2 Jane Manzi, PhD,1 Lesley A. Stevens, MD,3 Frederick Van Lente, PhD,4 David A. Lacher, MD, MEd,5 Andrew S. Levey, MD,3 and Josef Coresh, MD, PhD1,2

Prevalence of Chronic Kidney Disease in the United States

Josef Coresh, MD, PhD
Elizabeth Selvin, PhD, MPH
Lesley A. Stevens, MD, MS
Jane Manzi, PhD
John W. Kusek, PhD
Paul Eggers, PhD
Frederick Van Lente, PhD
Andrew S. Levey, MD
ARIC Study: N=13,500 stored serum samples collected in 1990-1992

2 R01s (Lutsey/Selvin)– shared 1 ml serum at U of Minnesota Lab (Steffes / Eckfeldt)

<table>
<thead>
<tr>
<th>Analyte</th>
<th>Module</th>
<th>Volume (uL)</th>
<th>Dead Volume</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modular-P</td>
<td>Fructosamine</td>
<td>13</td>
<td></td>
<td>Modular-P</td>
</tr>
<tr>
<td></td>
<td>Glycated albumin (+ albumin)</td>
<td>6.5</td>
<td></td>
<td>Modular-P</td>
</tr>
<tr>
<td></td>
<td>1,5-anhydroglucitol</td>
<td>5</td>
<td></td>
<td>Modular-P</td>
</tr>
<tr>
<td></td>
<td>Phosphorous</td>
<td>10</td>
<td></td>
<td>Modular-P</td>
</tr>
<tr>
<td></td>
<td>Calcium</td>
<td>7</td>
<td></td>
<td>Modular-P</td>
</tr>
<tr>
<td></td>
<td>GGT/ALT/AST</td>
<td>20</td>
<td></td>
<td>Modular-P</td>
</tr>
<tr>
<td></td>
<td>Beta-2-microglobulin</td>
<td>3</td>
<td></td>
<td>Modular-P</td>
</tr>
<tr>
<td></td>
<td>Cystatin-C</td>
<td>2</td>
<td></td>
<td>Modular-P</td>
</tr>
<tr>
<td></td>
<td>hs-CRP</td>
<td>10</td>
<td></td>
<td>Modular-P</td>
</tr>
<tr>
<td><strong>Analyte Volume</strong></td>
<td></td>
<td><strong>76.5</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>FGF23</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Analyte Volume</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dead Volume</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Volume</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELISA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Analyte Volume</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dead Volume</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Volume</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>LC Mass Spec</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Analyte Volume</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dead Volume</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Volume</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Elecsys</td>
<td>TSH</td>
<td>50</td>
<td></td>
<td>Elecsys</td>
</tr>
<tr>
<td></td>
<td>Free-T4</td>
<td>15</td>
<td></td>
<td>Elecsys</td>
</tr>
<tr>
<td></td>
<td>T3</td>
<td>15</td>
<td></td>
<td>Elecsys</td>
</tr>
<tr>
<td></td>
<td>TPOAb</td>
<td>20</td>
<td></td>
<td>Elecsys</td>
</tr>
<tr>
<td></td>
<td>PTH</td>
<td>50</td>
<td></td>
<td>Elecsys</td>
</tr>
<tr>
<td></td>
<td>hs-troponin T</td>
<td>15</td>
<td></td>
<td>Elecsys</td>
</tr>
<tr>
<td></td>
<td>NT-proBNP</td>
<td>15</td>
<td></td>
<td>Elecsys</td>
</tr>
<tr>
<td><strong>Analyte Volume</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dead Volume</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Volume</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Analyte Volume Sum**

Dead volume sum: 536.5
Total volume: 771.5

22 tests in <1 ml
tremendous coordination;
hard work

Epidemiologic analysis of the associations of these biomarkers with clinical outcomes during subsequent >20 years of follow-up of ARIC participants are in progress.
Fructosamine and glycated albumin for risk stratification and prediction of incident diabetes and microvascular complications: a prospective cohort analysis of the Atherosclerosis Risk in Communities (ARIC) study

Elizabeth Selvin, Andreea M Rawlings, Morgan Grams, Ronald Klein, A Richey Sharrett, Michael Steffes, Josef Coresh

Journal of the American Heart Association
Fibroblast Growth Factor-23 and Incident Coronary Heart Disease, Heart Failure, and Cardiovascular Mortality: The Atherosclerosis Risk In Communities Study

Pamela L. Lutsey, PhD, MPH; Alvaro Alonso, MD, PhD; Elizabeth Selvin, PhD, MPH; James S. Pankow, PhD, MPH; Erin D. Michos, MD, MHS; Sunil K. Agarwal, MD, PhD, MPH; Laura R. Loehr, MD, PhD, MS; John H. Eckfeldt, MD, PhD; Josef Coresh, MD, PhD, MHS
Minimizing error

- Implement a rigorous and detailed protocol, use standardized procedures and measurements
- Use reliable and accurate instruments and methods
- Rigorously train and re-train staff
- Conduct measurements consistently and correctly
- Have data checks in place to identify and prevent errors in data recording and entry
- Conduct continuous QC and QA analyses to identify problems and address them
- Obtain fasting blood samples (e.g. for glucose, cholesterol)
- Conduct repeated measurements if possible
- Build in reliability and validity studies into your protocol
Advice for stored specimen studies

- Build in reliability and validity studies whenever possible
  - Repeat measurements, include blind duplicate samples, evaluate freeze-thaw
  - Do method comparison studies, calibration

- Maximize use of existing data and valuable stored biospecimens
  - Efficiency of working with existing well-characterized cohorts
  - Work with colleagues to measure many things in a single biospecimen

- Consider partnering with industry
  - Don’t be afraid to ask for free stuff
  - Expand science, make budget go further
THANK YOU!