Measurement error and missing data
Killing two birds with one stone

Ruth Keogh
Department of Medical Statistics
London School of Hygiene & Tropical Medicine

On Behalf of TG4 Measurement Error and Misclassification
What are people saying and doing about measurement error?

We surveyed the literature in four areas:

• Nutritional intake cohort studies
• Physical activity cohort studies
• Air pollution cohort studies
• Dietary intake distributions
What percentage of studies mentioned measurement error as a potential problem?

What percentage of studies used methods to mitigate the impact of measurement error?

What percentage of studies categorized their main exposure?
What percentage of studies mentioned measurement error as a potential problem?  

80% (N=65)

What percentage of studies used methods to mitigate the impact of measurement error?

What percentage of studies categorized their main exposure?
Literature survey: N=81

What percentage of studies mentioned measurement error as a potential problem?

80% (N=65)

What percentage of studies used methods to mitigate the impact of measurement error?

6% (N=5)

What percentage of studies categorized their main exposure?
What percentage of studies mentioned measurement error as a potential problem? 80% (N=65)

What percentage of studies used methods to mitigate the impact of measurement error? 6% (N=5)

What percentage of studies categorized their main exposure? 88% (N=71)
• Most of those who mentioned error as a problem made an incomplete/incorrect claim
  – Many stated that their estimates could only be attenuated by measurement error
  – Some claimed no bias in associations but for spurious reasons
Literature survey: observations

• Most of those who mentioned error as a problem made an incomplete/incorrect claim
  – Many stated that their estimates could only be attenuated by measurement error
  – Some claimed no bias in associations but for spurious reasons

• Most studies categorized the continuous exposures
  – Common belief: categorization will reduce impact of measurement error
  – Categorizing can actually make things worse
Epidemiologic analyses with error-prone exposures: review of current practice and recommendations.


Measurement error is often neglected in medical literature: a systematic review.


Five myths about measurement error in epidemiologic research.

van Smeden, Lash, Groenwold. DOI 10.17605/OSF.IO/MSX8D. https://osf.io/msx8d/
Forthcoming guidance papers

STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology

Part 1 – basic theory and simple methods of adjustment
Ruth H Keogh, Pamela A Shaw, Paul Gustafson, Raymond J Carroll, Veronika Deffner, Kevin W Dodd, Helmut Küchenhoff, Janet A Tooze, Michael P Wallace, Victor Kipnis, Laurence S Freedman

Part 2 – more complex methods of adjustment and advanced topics
Pamela A Shaw, Paul Gustafson, Raymond J Carroll, Veronika Deffner, Kevin W Dodd, Ruth H Keogh, Victor Kipnis, Janet A Tooze, Michael P Wallace, Helmut Küchenhoff, Laurence S Freedman
Measurement error and missing data
Killing two birds with one stone

Ruth Keogh
Department of Medical Statistics
London School of Hygiene & Tropical Medicine

On Behalf of TG4 Measurement Error and Misclassification
Jonathan Bartlett, University of Bath

Christen Gray, IQVIA
Notation and set-up

True outcome $Y$

Exposure $X$

Confounder $Z$
Notation and set-up

$$X^* = X + \epsilon$$
True outcome model: Using $X$

$$X^* = X + \epsilon$$

$$Y = \beta_0 + \beta_X X + \beta_Z Z + \epsilon$$
Notation and set-up

True outcome model: Using $X$

$Y = \beta_0 + \beta_X X + \beta_Z Z + e$

Naive outcome model: Using $X^*$

$Y = \beta_0^* + \beta_X^* X^* + \beta_Z^* Z + e$
To do something about the impact of measurement error in our analysis, we need to know the form and extent of the error.
Motivating example: NHANES data

- Association between systolic blood pressure (SBP) and deaths due to cardiovascular disease (CVD)
- Adjusted for sex, age, smoking status, diabetes
- Analysis method: Cox regression
Motivating example: NHANES data

• Association between systolic blood pressure (SBP) and deaths due to cardiovascular disease (CVD)
• Adjusted for sex, age, smoking status, diabetes
• Analysis method: Cox regression

Challenges

• SBP is error-prone
• Missing data in smoking status
Motivating example: NHANES data

• Association between systolic blood pressure (SBP) and deaths due to cardiovascular disease (CVD)
• Adjusted for sex, age, smoking status, diabetes
• Analysis method: Cox regression

Challenges

• SBP is error-prone
• Missing data in smoking status

N=6519
Smoking observed
N=2667
5%
Replicate SBP measurement
Regression calibration

Obtain an estimate of $E(X|X^*, Z)$ using the ancillary study and use in the outcome regression model:

$$Y = \beta_0 + \beta_X E(X|X^*, Z) + \beta_Z Z + e$$
Regression calibration

Obtain an estimate of $E(X|X^*, Z)$ using the ancillary study and use in the outcome regression model:

$$Y = \beta_0 + \beta_X E(X|X^*, Z) + \beta_Z Z + e$$

Limitations

• Requires non-differential error assumption
• Requires an approximation for non-linear outcome models
• How do we accommodate missing data as well?
Multiple imputation (MI)

- Very popular method for handling missing data
- Measurement error can be viewed as a missing data problem – the ‘truth’ is missing
Multiple imputation (MI)

- Very popular method for handling missing data
- Measurement error can be viewed as a missing data problem – the ‘truth’ is missing

...for some people

Validation study

Replicates study

Calibration study

...for everyone!

1. For individuals with $X$ missing, draw a value $X$ from $X|X^*, Z, Y$
2. This gives a complete imputed data set
3. Fit the outcome model using the imputed data
4. Repeat for $M$ imputed data sets
5. Pool the results using Rubin’s Rules
Multiple imputation (MI)

In the validation situation we benefit from the huge missing data literature on MI.

**Carpenter & Kenward.** Multiple imputation and its application. New York: Wiley. 2013

**Sterne et al.** Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009; 338: b2393
Multiple imputation (MI)

In the validation situation we benefit from the huge missing data literature on MI.

**Carpenter & Kenward.** Multiple imputation and its application. New York: Wiley. 2013

**Sterne et al.** Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009; 338: b2393

**Software**

R: mice, smcfcs
Stata: mi impute, smcfcs
SAS: PROC MI
Multiple imputation (MI)


**Keogh & White.** A toolkit for measurement error correction, with a focus on nutritional epidemiology. Stat Med 2014; 33: 2137-2155.
Multiple imputation (MI)

The difficult step of MI

1. For individuals with $X$ missing, draw a value $X$ from $X|X_1^*, X_2^*, Z, Y$

• We need to e.g. assume a multivariate normal distribution for $X, X_1^*, X_2^*|Z$
• This gives form of $p(X|X_1^*, X_2^*, Z, Y)$

Replicates study
The difficult step of MI

1. For individuals with $X$ missing, draw a value $X$ from $X | X_1^*, X_2^*, Z, Y$

- We need to assume a distribution for $X, X_1^*, X_2^* | Z$, e.g. multivariate normal
- This gives form of $p(X | X_1^*, X_2^*, Z, Y)$

- This approach is not very flexible
- There is no software and it is not very easy to implement
In general it is difficult to know what is the form of $X|X_1^*, X_2^*, Z, Y$

- There are non-linear terms in the model
  \[ Y = \beta_0 + \beta_X X + \beta_Z Z + \beta_{X^2} X^2 + e \]

- The outcome model is not a linear regression
  \[ h(t|X, Z) = h_0(t) e^{\beta_0 + \beta_X X + \beta_Z Z} \]
A more flexible MI approach

In general it is difficult to know what is the form of $X|X_1^*, X_2^*, Z, Y$

- There are non-linear terms in the model
  
  $$Y = \beta_0 + \beta_x X + \beta_z Z + \beta_{x2} X^2 + e$$

- The outcome model is not a linear regression
  
  $$h(t|X, Z) = h_0(t)e^{\beta_0 + \beta_x x + \beta_z z}$$


A more flexible MI approach

Instead of trying to specify $X|X_{1^*}, X_{2^*}, Z, Y,$

...we specify $Y|X, Z$ and $X|Z$ and the measurement error model

Basic idea
1. Propose a potential imputed value for $X$ from $X|X_{1^*}, X_{2^*}, Z$
2. Use a rejection sampling procedure to accept or reject the value as being from the target distribution $X|X_{1^*}, X_{2^*}, Z, Y$
3. The acceptance/rejection rule is a function of the outcome model

Substantive model compatible full conditional specification (SMCFCS)
A more flexible MI approach

**Application for measurement error correction**

- Validation study: we can use it directly
- Replicates: we extended the method to the setting of replicates


**Bartlett & Keogh.** smcfcs: Multiple imputation of covariates by substantive model compatible fully conditional specification. 2019.

https://github.com/ruthkeogh/meas_error_handbook
Motivating example: NHANES data

- Association between systolic blood pressure (SBP) and deaths due to cardiovascular disease (CVD)
- Adjusted for sex, age, smoking status, diabetes
- Analysis method: Cox regression

Challenges

- SBP is error-prone
- Missing data in smoking status

Replicate SBP measurement

N=6519
Smoking observed
N=2667
5%
Motivating example: NHANES data

First ignoring missing data.....

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Naïve analysis</th>
<th>Regression calibration</th>
<th>Multiple imputation</th>
</tr>
</thead>
<tbody>
<tr>
<td>SBP</td>
<td>0.085 (0.014, 0.157)</td>
<td>0.114 (0.011, 0.222)</td>
<td>0.120 (0.020, 0.219)</td>
</tr>
<tr>
<td>Male</td>
<td>0.49 (0.30, 0.67)</td>
<td>0.49 (0.32, 0.68)</td>
<td>0.49 (0.30, 0.67)</td>
</tr>
<tr>
<td>Age</td>
<td>0.88 (0.77, 0.99)</td>
<td>0.87 (0.76, 0.99)</td>
<td>0.88 (0.77, 0.99)</td>
</tr>
<tr>
<td>Smoker</td>
<td>0.26 (0.07, 0.46)</td>
<td>0.26 (0.07, 0.45)</td>
<td>0.26 (0.07, 0.46)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>0.50 (0.29, 0.72)</td>
<td>0.50 (0.28, 0.72)</td>
<td>0.50 (0.29, 0.72)</td>
</tr>
</tbody>
</table>
Motivating example: NHANES data

Accounting for missing data as well...

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Naïve analysis</th>
<th>Regression calibration</th>
<th>Multiple imputation</th>
<th>Multiple imputation 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>SBP</td>
<td>0.085 (0.014, 0.157)</td>
<td>0.114 (0.011, 0.222)</td>
<td>0.120 (0.020, 0.219)</td>
<td>0.104 (0.035, 0.173)</td>
</tr>
<tr>
<td>Male</td>
<td>0.49 (0.30, 0.67)</td>
<td>0.49 (0.32, 0.68)</td>
<td>0.49 (0.30, 0.67)</td>
<td>0.46 (0.35, 0.56)</td>
</tr>
<tr>
<td>Age</td>
<td>0.88 (0.77, 0.99)</td>
<td>0.87 (0.76, 0.99)</td>
<td>0.88 (0.77, 0.99)</td>
<td>1.04 (0.97, 1.11)</td>
</tr>
<tr>
<td>Smoker</td>
<td>0.26 (0.07, 0.46)</td>
<td>0.26 (0.07, 0.45)</td>
<td>0.26 (0.07, 0.46)</td>
<td>0.26 (0.09, 0.43)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>0.50 (0.29, 0.72)</td>
<td>0.50 (0.28, 0.72)</td>
<td>0.50 (0.29, 0.72)</td>
<td>0.69 (0.56, 0.83)</td>
</tr>
</tbody>
</table>

N=2667
N=6519
• We commonly face more than one ‘data quality’ challenge at the same time

• Multiple imputation (and fully Bayesian approaches) enable us to ‘easily’ tackle measurement error and missing data together

• The smcfcs package in R facilitates this
Summary

• We commonly face more than one ‘data quality’ challenge at the same time

• Multiple imputation (and fully Bayesian approaches) enable us to ‘easily’ tackle measurement error and missing data together

• The smcfc package in R facilitates this
