Understanding and adjusting for Berkson error arising from prediction equations

Pamela Shaw

STRATOS TG4 Measurement error and Misclassification

Pamela.A.Shaw@kp.org

STRATOS Meeting

November 2, 2021

KAISER PERMANENTE®

Acknowledgments

This is joint work with a subgroup of members of **STRATOS TG4: Measurement Error and Misclassification Topic Group** and collaborators

- Laurence Freedman (Gertner Institute, STRATOS TG4 chair)
- Victor Kipnis (NCI, STRATOS TG4 chair)
- Paul Gustafson (UBC, STRATOS TG4 member)
- Doug Midthune (NCI, STRATOS TG4 member)
- Daniela Sotres-Alvarez (UNC-Chapel Hill)
- Lillian Boe (UPenn)
- Jenny Shen (UPenn)
- Eunyoung Park (UPenn)

Introduction

- In epidemiology, there are many measurements that are difficult to obtain directly:
  - Expensive (Resting Energy Expenditure)
  - Burdensome (24-hour urinary sodium)
  - Impossible (Usual energy intake)

- One strategy is to use prediction equations to measure them indirectly

- Many analyses proceed with predicted values as if they were observed data

- Using predicted values instead of observed data in study analyses can corrupt study results if the (Berkson) prediction error is not handled appropriately

Planning a series of papers examining issues that arise when predicted values are used in data analysis:

- Paper 1: Introductory concepts + Example of estimating of a distribution
- Paper 2: Analytical issues that arise when applying regression calibration

Paper 1: Introduction + Estimation of a distribution

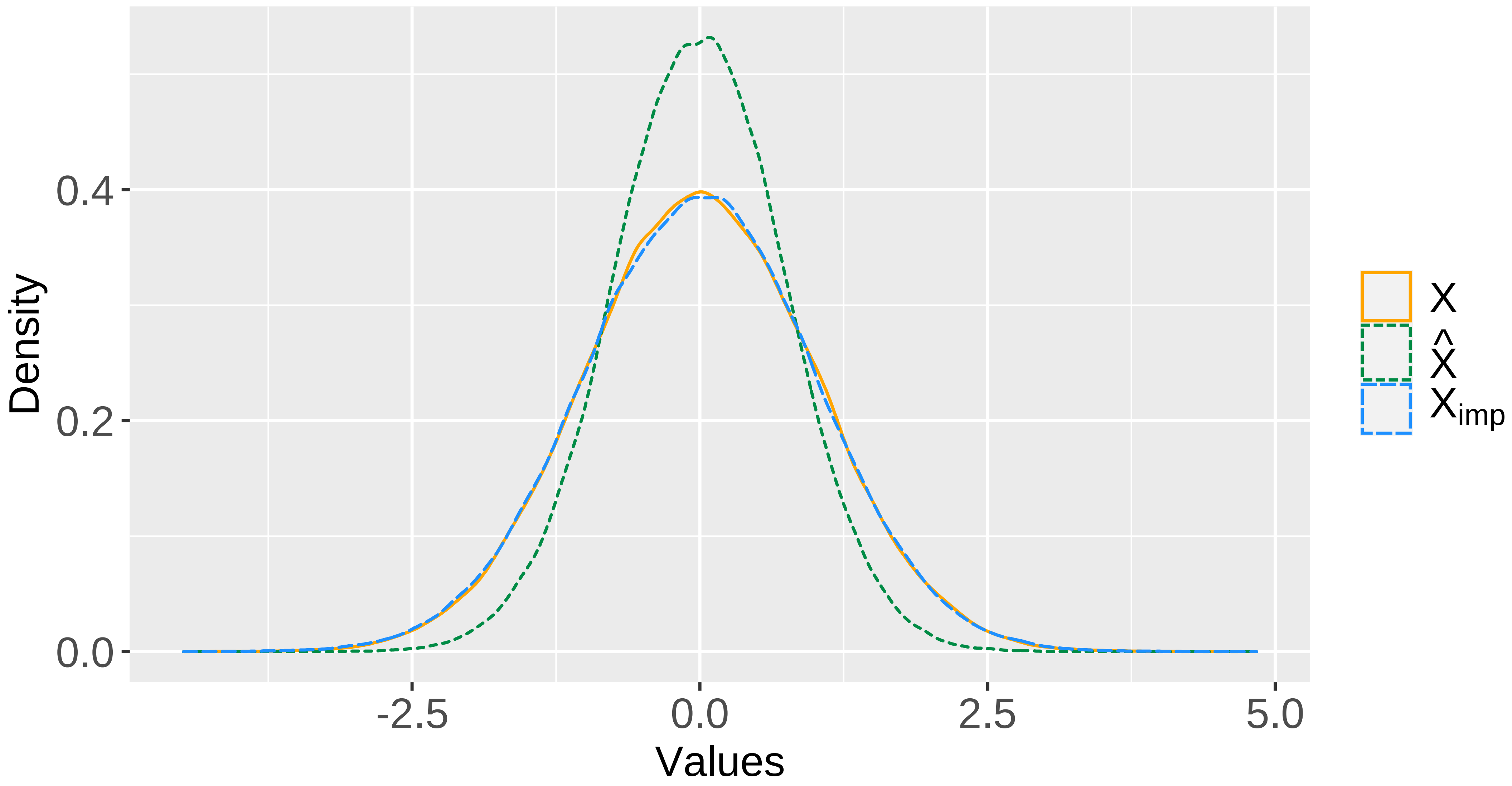
- Consider setting where have an error prone X* and use a predicted value $\hat{X}$ to correct for systematic and random error
- Introduction to prediction error as Berkson measurement error

  $X = \hat{X} + \text{error}$
- Examine effects of ignoring prediction/Berkson error when estimating a distribution
- Present a simple, novel method to handle Berkson error in this setting
- Concepts illustrated with simulated data where truth is known
- Data example from a complex survey design

A simple fix for Berkson error

- **A fundamental attribute of predicted values is their Berkson error makes them less variable than they should be**
- **A simple fix is to add back the missing variance to the calibrated value.**
  - This can be accomplished from simulating error e ~ $(0,\sigma^2)$
  - $X_{imp} = \hat{X} + e$
  - A multiple imputation approach is applied to estimate quantities (Baldoni et al 2021)
    - Applied in the context of a complex survey design

Simulation study results



Berkson error biases quantiles and standard errors

	X			$\hat{X}$			X_{imp}		
%-tile	Mean	ESE	CP	Mean	ESE	CP	Mean	ESE	CP
25th	-0.672	0.043	94.8	-0.501	0.079	8.0	-0.679	0.089	96.6
50th	-0.001	0.039	96.1	-0.002	0.067	6.0	-0.002	0.074	97.4
75th	0.674	0.043	94.5	0.498	0.078	8.3	0.675	0.087	96.5

Example from the Hispanic Community Health Study

(Lavange et al 2010)

Question of interest: Does sodium intake vary by Hispanic ethnicity?

HCHS main cohort: n = 16,415 (Chicago, Miami, New York, San Diego)

Male: 40%

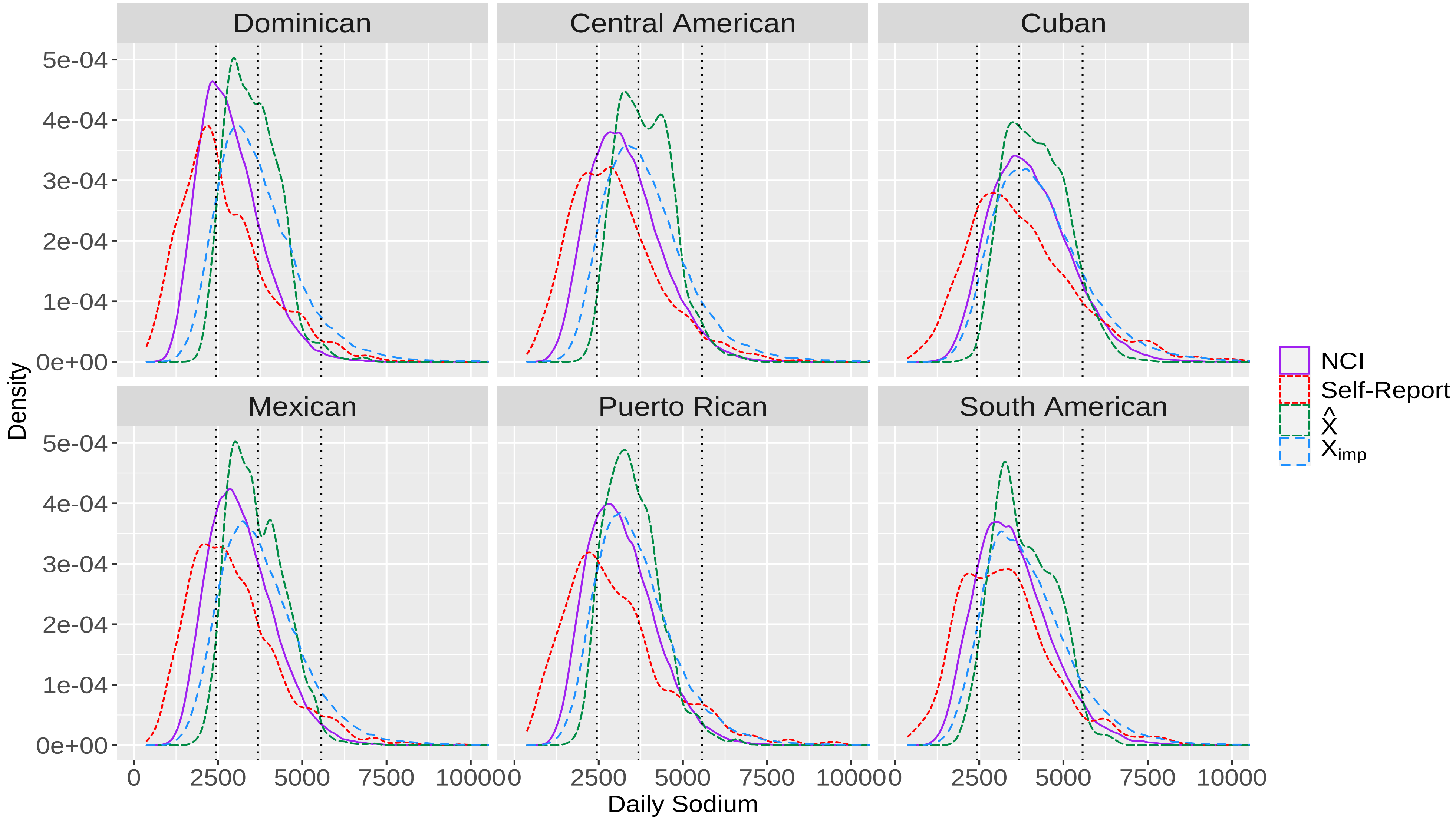
Age: mean 43y; range: 18-74y

Main dietary assessment X*: two 24-hour recalls, known to be subject to bias

SOLNAS: Calibration sub-study: n = 477

Biomarkers X**: Doubly-labeled water (energy) and 24-hour urinary markers (protein, potassium, sodium) were obtained to create calibration equations that correct for the measurement error/bias in self-reported sodium (Mossavar-Rahmani et al 2017)

Similar results seen in HCHS/SOL



Paper 2: Analytical issues that arise when applying regression calibration (RC)

- RC is the most common method to address covariate measurement error
- RC involves replacing unobserved error-free covariate X with a predicted value in outcome model (e.g. $\hat{X}$=E[X|X*,Z])
- Analytical issues generalize setting with a predicted covariate in a regression model

Considerations for analysis

Regression calibration relies on:

- All the covariates in the outcome model to be in the calibration model
- Prediction error independent of the outcome
- Adjustment to the standard error calculation to account for extra uncertainty
  - The usual standard errors from regression software are too small
  - The bootstrap or sandwich estimators are two options.
- Interesting analytical issues arise if there is a mediator in the model

Regression calibration and mediation: a dilemma

- Generally, if you are interested in the total effect of X on Y then you should not include M in the outcome model
- If M is an important variable in the calibration model, it should generally be included in the outcome model to avoid bias when applying regression calibration

Example:

BMI is one of the strongest predictors of energy intake and may mediate the effect of energy intake on outcomes like heart disease, cancer, diabetes

Mediation

Some notation

- Y = outcome variable
- X = exposure of interest
- Z = confounder(s)
- M = mediator

The models

- $M = \gamma_0 + \gamma_X X + \gamma_Z Z + \delta,$ (1) Mediation model
- $Y = \beta_0 + \beta_X X + \beta_Z Z + \beta_M M + \varepsilon,$ (2) Outcome model

Substituting the right-hand side of equation (1) for M in equation (2), we get

- $Y = \tilde{\beta}_0 + \tilde{\beta}_X X + \tilde{\beta}_Z Z + \tilde{\varepsilon},$ **where $\tilde{\boldsymbol{\beta}}_X = \boldsymbol{\beta}_X + \boldsymbol{\beta}_M \boldsymbol{\gamma}_X$**

Where β_X is the direct effect, and $\beta_M \gamma_X$ is the indirect effect
(this method is approximate for non-linear models)

Addressing Mediation with Regression Calibration

The Models:

- $M = \gamma_0 + \gamma_X X + \gamma_Z Z + \delta,$ (1) Mediation model
- $Y = \beta_0 + \beta_X X + \beta_Z Z + \beta_M M + \varepsilon,$ (2) Outcome model

Midthune Method (Freedman et al (2011))

Step 1 Estimate γ_X from equation (1) using RC to adjust for ME

Step 2: estimate β_X and β_Z from equation (2) using RC to adjust for ME

Step 3: Estimate $\tilde{\beta}_X$ using the equation $\tilde{\beta}_X = \beta_X + \beta_M \gamma_X$.

Mediation Results from HCHS/SOL

Binary Outcome Y: High risk for metabolic syndrome
Exposure of interest X: Energy Intake – estimated OR for 20% increase
Mediator M: BMI
X*: self-reported intake using 24 hour recalls
Z: age, Hispanic/Latino background, education, income, and current smoking.

Method of Estimation	OR	95% CI
Including BMI in outcome model	0.85	0.46 – 1.58
Omitting BMI from outcome model	3.76	3.03 – 4.67
Midthune's method	1.52	1.02 – 2.25

Discussion

- There is increasing use of prediction and calibration equations in medicine
- Naïve analyses with predicted outcomes are subject to multiple biases
- Presented methods do not address when error is differential
- Awareness of the effects of Berkson error and methods to adjust for it need more attention

References

- Baldoni PL, Sotres-Alvarez D, Lumley T, Shaw PA. On the Use of Regression Calibration in a Complex Sampling Design With Application to the Hispanic Community Health Study/Study of Latinos. American Journal of Epidemiology. 2021 Jan 28.
- Buonaccorsi J. Measurement errors, linear calibration and inferences for means. *Comp stat and Data Analysis,*1991;11(3):239-57.
- Freedman LS, Midthune D, Carroll RJ, Tasevska N, Schatzkin A, Mares J, Tinker L, Potischman N, Kipnis V. Using regression calibration equations that combine self-reported intake and biomarker measures to obtain unbiased estimates and more powerful tests of dietary associations. Am J Epidemiol 2011; 174:1238-1245.
- Haber G, Sampson J, Graubard B. Bias due to Berkson error: issues when using predicted values in place of observed covariates. Biostatistics. 2020 Feb 10.
- Haber G, Sampson J, Flegal KM, Graubard B. The perils of using predicted values in place of observed covariates: an example of predicted values of body composition and mortality risk. The American Journal of Clinical Nutrition. 2021 Apr 8.
- Keogh RH, Shaw PA, Gustafson P, Carroll RJ, Deffner V, Dodd KW, Küchenhoff H, Tooze JA, Wallace MP, Kipnis V, Freedman LS. STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part I – basic theory, validation studies and simple methods of adjustment. *Statistics in Medicine* 2020 Jul 20;39(16):2197-2231.
- Lavange L et al. Sample design and cohort selection in the Hispanic Community Health Study/Study of Latinos. *Ann Epidemiol.* 2010;20(8):642-649.

References – Part 2

- Mossavar-Rahmani Y, Sotres-Alvarez D, Wong W, Loria C, Gellman M, Van Horn L, Alderman M, Beasley J, Lora C, Siega-Riz AM, Kaplan R, Shaw PA. Applying recovery biomarkers to calibrate self-report measures of sodium and potassium in the Hispanic Community Health Study/Study of Latinos Journal of Human Hypertension, 2017; 31(7): 462-473, Jul 2017.
- Ogburn EL, Rudolph KE, Morello-Frosch R, Khan A, Casey JA. A Warning About Using Predicted Values From Regression Models for Epidemiologic Inquiry. American Journal of Epidemiology, In Press
- Prentice RL, Huang Y, Kuller LH, et al. Biomarker-calibrated energy and protein consumption and cardiovascular disease risk among postmenopausal women. Epidemiology (Cambridge, Mass.). 2011 Mar;22(2):170.
- Prentice RL, Shaw PA, Bingham SA, et al. Biomarker-calibrated energy and protein consumption and increased cancer risk among postmenopausal women. American Journal of Epidemiology. 2009 Apr 15;169(8):977-89.
- Prentice RL. Covariate measurement errors and parameter estimation in a failure time regression model. Biometrika. 1982 Aug 1;69(2):331-42.
- Shaw PA, Gustafson P, Carroll RJ, Deffner V, Dodd KW, Keogh RH, Kipnis V, Tooze JA, Wallace MP, Küchenhoff H, Freedman LS. STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part II –more complex methods of adjustment and advanced topics. *Statistics in Medicine* 2020 Jul 20;39(16):2232-2263.
- Shaw PA, Deffner V, Dodd KW, Freedman LS, Keogh RH, Kipnis V, Kuechenhoff H, and Tooze JA for the Measurement Error and Misclassification Topic Group (TG4) of the STRATOS Initiative: Epidemiologic analyses with error-prone exposures: Review of current practice and recommendations. *Annals of Epidemiology,* 2018; 28(11): 821-828.
- Tooze JA, Kipnis V, Buckman DW et al. A mixed-effects model approach for estimating the distribution of usual intake of nutrients: the NCI method. Statistics in Medicine. 2010 Nov 30;29(27):2857-68.