Efficient Regression Calibration for Logistic Regression
in Main Study/Internal Validation Study Designs
with an Imperfect Reference Instrument

Donna Spiegelman¹, Raymond J. Carroll², Victor Kipnis³

February 22, 2000

Sources of support: NIH CA50597¹, NIH CA74112¹, NIH ES09411¹, NIH CA57030², P30-ES09106²

1 Departments of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, MA 02115. e-mail: stdls@channing.harvard.edu

2 Department of Statistics, Texas A&M University, College Station, Texas 77843. e-mail carroll@stat.tamu.edu

³NCI Biometry Research Group EPN-344, 6130 Executive Boulevard MSC 7354, Bethesda, MD 20892. e-mail vk3b@nih.gov

DRAFT: Please do not quote or circulate without authors’ permission
Efficient Regression Calibration for Logistic Regression in Main Study/Internal Validation Study Designs, with an Imperfect Reference Instrument

Abstract

An extension to the version of the regression calibration estimator proposed by Rosner et al. (1990) for logistic and other generalized linear regression models is given for main study/internal validation study designs. This estimator combines the information about the parameter of interest contained in the internal validation study with Rosner et al.’s regression calibration estimate, using a generalized inverse-variance weighted average. It is shown that the validation study selection model can be ignored as long as this model is jointly independent of the outcome and the incompletely observed covariates, conditional, at most, upon the surrogates and other completely observed covariates. In an extensive simulation study designed to follow a complex, multivariate setting in nutritional epidemiology, it is shown that with validation study sizes of 340 or more, this estimator appears to be asymptotically optimal in the sense that it is nearly unbiased and nearly as efficient as a properly specified maximum likelihood estimator. A modification to the regression calibration variance estimator which replaces the standard uncorrected logistic regression coefficient variance with the sandwich estimator to account for the possible mis-specification of the logistic regression fit to the surrogate covariates in the main study (Kuha, 1994), was also studied in this same simulation experiment. In this study, the alternative variance formula yielded results virtually identical to the original formula. A version of the proposed estimator is also derived for the case where the reference instrument, available only in the validation study, is imperfect but unbiased at the individual level and contains error that is uncorrelated with other covariates and with error in the surrogate instrument. Replicate measures are obtained in a subset of study participants. In this case, it is shown that the validation study selection model can be ignored when sampling into the validation study depends, at most, only upon perfectly measured covariates. Two data sets, a study of fever in relation to occupational exposure to antineoplastics among hospital pharmacists and a study of breast cancer incidence in relation to dietary intakes of alcohol and vitamin A, adjusted for total energy intake, from the Nurses’ Health Study, were analyzed using these new methods. In these data, because the validation studies contained less than 200 observations and the events of interest were relatively rare, as is typical, the potential improvements offered by this new estimator were not apparent.
1 Introduction

Risk factors studied in epidemiology are often subject to measurement error. If one variable is measured with error, it is often but not always the case that point estimates of its effect will be under-estimated. However, if more than one variable is included in a model and one or more of these are measured with error, the estimates of any of the effects may be under- or over- estimated, even those corresponding to model covariates measured without error. There has been considerable interest in the effects of covariate measurement error in recent years and many papers have been written proposing methods to adjust for these errors. Recent non-technical reviews of this extensive statistical and epidemiologic literature have appeared, as has a comprehensive review.

When a gold standard is available, a validation study can be conducted in which the usual, error-prone exposure measure ($X$) is validated against the gold standard ($x$). If more than one variable is measured with error, these variables can be jointly validated against their gold standards within the same sub-sample of study participants, permitting control for any correlations in their errors. An internal validation study is available when the study subjects who contribute validation data are a subsample of the main study. In this case, the status of the outcome variable is known for each validation study participant. Otherwise, the validation study is external. Validation data are typically far more costly to obtain than data from the main study, and guidelines for cost-efficient main study/validation study designs have been given. Methods proposed in the literature to handle the problem of covariate measurement error use such validation data, along with the usual main study data, to obtain consistent point and interval estimates of exposure effects.

The non-iterative regression calibration method can be used to obtain (approximately) consistent point estimates and valid interval estimates of relative risk in regression models with measurement error in one or more continuous covariates. The validation study
is used to estimate the regression model for $E(x/X)$. True covariates are predicted for all main study subjects using this model, and the regression of the outcome of interest on the estimated $x$’s is then run in the main study$^4$. In another version of this procedure, originally developed by Rosner et al. for multiple logistic regression, the uncorrected point and interval estimates of exposure effects from the main study are explicitly corrected using the linear regression model for $E(x/X)$ estimated from the validation data$^8$.$^9$. The key assumption for valid use of regression calibration is that measurement error is non-differential with respect to the response variable, $Y$, i.e. $f(Y|x,X)=f(Y|X)$. The regression calibration approach does not require that a gold standard be available in the validation study; it can also be applied when an imperfect reference instrument, $x^*$, is available, as long as the errors in $x^*$ are uncorrelated with the true covariates $x$ and the errors in $X$. Examples using this method for reporting results in original scientific publication have been published $^{13}$ $^{14}$ $^{15}$ $^{16}$ $^{17}$ $^{18}$ $^{19}$.

In this paper, we extend the formulation of the original methodology to make more efficient use of the available data when the validation study contains outcome data, i.e. when internal validation data are available. A further extension provides approximately unbiased point and interval estimates when an imperfect, unbiased reference instrument, $x^*$, also known as an unbiased alloyed gold standard$^{20}$, is available in the validation study, in addition to the outcome data, as long as the errors in $x^*$ are uncorrelated with $x$ and the errors in $X$. Although the proposed methodology can be used with a general class of regression models$^{12}$, we will concentrate on the case of logistic regression. In this case, we also investigate whether the use of the robust variance formula proposed by Kuha$^{21}$ appreciably improves the statistical properties of this method.

2 Methods
2.1 Review of Rosner et al.’s regression calibration (RC) method

The parameter of interest is $\beta_1$ from the generalized linear model

$$g[E(Y|x)] = \beta_0 + \beta_1 x,$$

where $g(\cdot)$ is a link function which linearizes the conditional mean function. Substituting the covariate measured with error, $X$, for $x$, the uncorrected point and interval estimates of the exposure effect, $\hat{\beta}_1$, obtained from fitting (1) are adjusted for measurement error in a simple one-step procedure. When measurement error is present in $X$, the estimated regression coefficient, $\hat{\beta}_1$, is biased for $\beta_1$, often severely so. When $g(\cdot) = E(Y/X)$, then the regression calibration is applied to a linear regression model \(^{10}\). When $g(\cdot) = \text{logit}[E(Y/X)]$, then the regression calibration is applied to a logistic regression model \(^{8,9}\). When $g[E(Y|X)] = [\log[I(t|X=0)] + \beta_1 X$, where $I(t)$ is the incidence rate at time $t$ and $g[E(Y|X)] = log[I(t|X)]$, then regression calibration can be applied to a Cox proportional hazards regression model \(^{11,22}\). The application of regression calibration to these three basic models, all of which are used widely in epidemiology, was unified with a special focus on interval estimation and computing in SAS \(^{11}\).

When there is one covariate in the model and this covariate is measured with error, the point and interval estimates of effect can be corrected for measurement error using the simple formulas

$$\hat{\beta}_{RC} = \hat{\beta}_1 / \hat{\gamma}, \text{ and}$$

$$\text{Var}(\hat{\beta}_{RC}) = \text{Var}(\hat{\beta}_1) / \hat{\gamma}^2 + \hat{\beta}_1^2 \text{Var}(\hat{\gamma}) / \hat{\gamma}^4$$

where $\hat{\beta}_1$ and $\text{Var}(\hat{\beta}_1)$ are obtained from fitting (1) to the main study data with $X$ substituted for $x$, and $\hat{\gamma}$ and $\text{Var}(\hat{\gamma})$ are obtained from fitting to the validation data the linear regression
where $x$ is the correctly measured exposure variable, and

$$Var(x|X) = \sigma^2.$$  \hspace{1cm} (4)

Regression calibration generalizes in a straightforward manner to the more realistic case when there is more than one covariate in the model, some of which are measured with error and some not\textsuperscript{8,9,10,11,12,13}.

The regression calibration estimator for logistic regression was initially derived by Rosner et al. assuming a linear homoscedastic regression of $x$ on $X$ with a normally distributed error term and using a mathematical approximation which required a rare disease\textsuperscript{23}. Although considering a broader class of estimators, Carroll and Stefanski\textsuperscript{24} showed that the assumption of normality of $f(x|X)$ is not needed; their version of the RC estimator algebraically identical to (2) can be derived for logistic regression making only assumptions (1), (3), and (4), which involve linearity of the probability of occurrence of disease on the logit scale, given the gold standard for variables measured with error and other perfectly measured covariates; linearity of the conditional mean of the gold standard given the usual exposure measurements and other perfectly measured covariates; homoscedasticity of the measurement error model variance; and small measurement error. Subsequently, Kuha showed that the key requirement for approximate unbiasedness of $\hat{\beta}_{RC}$ in the logistic regression setting is that either 1) $\beta_1^2 \sigma^2$ is small, or 2) $Pr(Y=1|x)$ is small and $f(x|X)$ is normal\textsuperscript{21}.

In the same paper, Kuha proposed that the robust variance of $\hat{\beta}$ be used rather than the 'naive' variance obtained from fitting the primary regression model of $Y$ on $X$ and $u$,

$$logit[E(Y|X,u)] = \beta_0 + X\beta_1 + u\beta_2$$  \hspace{1cm} (5)
where $X$ is now a row vector of variables measured with error, $u$ is a row vector of perfectly measured variables, and $\beta=(\beta_0, \beta_1, \beta_2)'$. Since $X$ is different from $x$, model (5) will, in general be mis-specified, and the results of Huber\textsuperscript{25} would apply, giving the estimated asymptotic covariance matrix for $\hat{\beta}$ as

$$\text{Var}_R(\hat{\beta}) = \frac{1}{n_1} I(\hat{\beta})^{-1} G(\hat{\beta}) I(\hat{\beta})^{-1} \quad (6)$$

where $I(\hat{\beta})^{-1}$ is the estimated uncorrected covariance matrix from the regression of $Y$ on $X$ and $u$, and $G(\hat{\beta})$ is the average over all $n_1$ main study subjects of $g_i g_i'$, where $g_i$ is main study subject $i$'s vector of quasi-score functions, evaluated at $\hat{\beta}$. Here, a quasi-score function is the derivative of the log of the Bernoulli likelihood implied by (3). In Section 4, we describe a simulation study of the properties of Wald-based inference about $\beta$, substituting the robust variance estimate for $\hat{\beta}$ for the standard variance estimate in (2) and its multivariate counterpart, compared with inference based upon the standard variance formula.

### 2.2 Extension for main study/internal validation study designs

#### 2.2.1 Gold standard available in the validation study

In a main study/internal validation study design, the primary regression coefficients, $\beta$, can be estimated without any measurement error bias in the validation study, given that all assumptions previously specified, including that $x$ is measured without error, are met. Consider, for example, the case of estimating a scalar slope $\beta_1$. We propose here a simple extension of the RC estimator which uses the internal validation data more efficiently through an inverse variance weighted average of the two estimates, i.e.
\( \hat{\beta}_{RC,I} = w_{RC} \hat{\beta}_{RC} + w_I \hat{\beta}_I \), where \( \hat{\beta}_{RC} \) is the usual RC estimator as given by (3), \( \hat{\beta}_I \) is the slope estimate obtained from the validation study data alone from the primary regression model \( \Phi \), \( w_{RC} = \text{Var}(\hat{\beta}_{RC})^{-1} \left[ \text{Var}(\hat{\beta}_{RC})^{-1} + \text{Var}(\hat{\beta}_I)^{-1} \right]^{-1} \), \( w_I = \text{diag} \), \( \text{Var}(\hat{\beta}_{RC}) \) is given by \( \Phi \) and \( \text{Var}(\hat{\beta}_I) \) is obtained by fitting model (1) in the validation study. As long as \( \hat{\beta}_I \) and \( \hat{\beta}_{RC} \) are asymptotically uncorrelated, as we show that they are in Appendix 1, this choice of weights gives us the asymptotically most efficient combined estimate among all unbiased linear combinations of \( \hat{\beta}_I \) and \( \hat{\beta}_{RC} \). The asymptotic variance of this estimator, \( \hat{\beta}_{RC,I} \), is given approximately by

\[
\text{Var}(\hat{\beta}_{RC,I}) = \left[ \text{Var}^{-1}(\hat{\beta}_{RC}) + \text{Var}^{-1}(\hat{\beta}_I) \right]^{-1}.
\]

This extension should result in a noticeable improvement in efficiency over the original RC estimator whenever the validation study is large enough to permit a reasonable estimate of \( \beta \). For example, this may happen in logistic regression analysis of cross-sectional studies where symptom prevalence is a common outcome, in cohort studies where the cumulative incidence of the outcome of interest is large, or when the dependent variable is continuous.

The multivariate version of \( \hat{\beta}_{RC,I} \) follows directly from the theory of generalized least squares. Under the assumption that model (5) is exact, for a suitable \( 2(p+q+1) \times (p+q+1) \) matrix of zeroes and ones, \( Z \), one can write

\[
\begin{pmatrix} \hat{\beta}_{RC} \\ \hat{\beta}_I \end{pmatrix}' - Z \beta + e
\]
where $E(e)=0$, $\text{dim}(x)=\text{dim}(X)=p$ and $\text{dim}(u)=q$. We show in Appendix 2 that as long as sampling into the validation study is unbiased,

$$\hat{\beta}_{RC,I} - (Z'\overline{V}^{-1}Z)^{-1}Z'\overline{V}^{-1}(\hat{\beta}_p') = -[\overline{V}_I^{-1} + \overline{V}_I^{-1}]^{-1}[\overline{V}_I^{-1}\hat{\beta}_I + \overline{V}_I^{-1}\hat{\beta}_I]$$

and

$$\text{Var}(\hat{\beta}_{RC,I}) = (Z'\overline{V}^{-1}Z)^{-1} - [\overline{V}_I^{-1} + \overline{V}_I^{-1}]^{-1},$$

where $\text{Var}(e) = \mathbb{V} = \text{diag}\{\overline{V}_I, \overline{V}_I\}$, $\overline{V}_RC$ is the covariance matrix estimate for $\hat{\beta}_{RC}$, and $\overline{V}_I$ is the covariance matrix estimate for $\hat{\beta}_I$. Assuming that (5) is exact, by generalized least squares theory $^{26}$, $\hat{\beta}_{RC,I}$ is unbiased for $\beta$ and minimum variance among all unbiased linear combinations of $\hat{\beta}_{RC}$ and $\hat{\beta}_I$, as long as the inverse variance weights are accurately estimated. An unbiased sampling mechanism is one in which the usual estimates are consistent without adjustment for sampling. It is shown in Appendix 2 that unbiased sampling will occur whenever sampling into the validation study is independent of $(Y,x)$ conditional upon $(X,u)$. Simple random sampling into the validation study will satisfy this condition, but it is not necessary.

### 2.2.2 Extension for main study/internal validation study designs with a replicated unbiased alloyed gold standard

In a main study/internal validation study design with an alloyed gold standard, there are $n_1$ main study subjects with data $(Y_i, X_i, u_i), i=1,...,n_1$, and $n_2$ validation study subjects with data $(Y_i, X_i, x_{i1}^*, ..., x_{iR}^*, u_i), i=n_1+1,...,n$, where $n=n_1+n_2$. For the $j^{th}$ measurement out of $n_R$ replicate measurements in the $i^{th}$ subject, an unbiased alloyed gold standard, $x_{ij}^*$, is one for which the model
\[ x_{ij}^* = x_i + e_{ij} \] (8)

applies, where \( \text{dim}(x_{ij}^*)=\text{dim}(e_{ij})=\text{dim}(x_i)=p \), \( \text{dim}(u_i)=q \), \( p+q \) equals the number of covariates in model (5), \( \text{Var}(e_{ij})=\Sigma \), \( \text{Var}(x_i)=\Sigma_x \), and we assume that \( E(e_{ij})=0 \), \( \text{Cov}(x_i,e_{ij})=0 \), and \( \text{Cov}(X_i,x_i,e_{ij})=0 \). For example, in the Nurses’ Health Study, nutrient intakes obtained by food frequency questionnaire were validated by daily weighed diet records obtained over four one-week periods. To the extent that nutrient intake of Vitamin A varies daily, even the average of 28 days of intake will not perfectly describe the subject’s average daily intake over the past year, and it may be reasonable to assume that these daily measurements follow model (8), although other models could be considered\(^{27}\). Then, as has been shown previously\(^{49}\), \( \hat{\beta}_1 \), the estimated log relative risk obtained from the logistic regression of \( Y \) on \((x^*_i,u_i)\) will be biased, approximately, by the factor \( R \), where \( R = \Sigma_{x_i,u_i}[\Sigma_{x_i,u_i}+\Sigma/n_R]^{-1} \), \( n_R \) is the number of replicate measurements of \( x_{ij}^* \) available for each validation study subject, \( x_i^* = n_R^{-1} \sum_{j=1}^{n_R} x_{ij}^* \), \( \Sigma_{x_i,u_i} = \text{Var}(x_i^*,u_i) \), with zeroes in the \( p+q+1 \)th row and column, \( \text{dim}(\Sigma_{x_i,u_i})=(p+q+1) \times (p+q+1) \), and \( \Sigma \) is augmented to \( \text{dim}(p+q+1) \times (p+q+1)\) with 0’s everywhere outside of the original \( p \times p \) block. For example, in the Nurses’ Health Study, \( n_R=26 \).

To obtain \( \hat{\beta}_{RCJ}\^* \) in this setting, the theory of generalized least squares is applied as previously. Under the assumption that model (5) is exact,
\[
(\hat{\beta}_{RC}', \hat{\beta}_I')' = Z\beta^* + \epsilon
\]

where \( \text{E}(\epsilon) = 0 \),

\[
\text{Var}(\epsilon) = V^* - 
\begin{bmatrix}
V_{RC} & V_{RC,I}^* \\
V_{RC,I}^* & V_I^*
\end{bmatrix}
\]

\( V_{RC} \) is the covariance matrix for \( \hat{\beta}_{RC} \), \( V_I^* \) is the covariance matrix for \( \hat{\beta}_I^* \), and \( V_{RC,I}^* \) is \( \text{Cov}(\hat{\beta}_{RC}, \hat{\beta}_I^*) \). Then, \( \hat{\beta}_{RC,I}^* = (Z'[V^*]^{-1}Z)^{-1}Z'[V^*]^{-1}\hat{\beta}^* \) and \( \text{Var}(\hat{\beta}_{RC,I}) = (Z'[V^*]^{-1}Z)^{-1} \), where \( \hat{\beta}^* \) is the vector on the left-hand side of (9). \( \text{Cov}(\hat{\beta}_{RC}, \hat{\beta}_I^*) \) is derived in Appendix 3 using generalized estimating equations theory, or can be estimated by a non-parametric bootstrap. As long as replicate data are available in the validation study and sampling into the validation study is unbiased, \( \hat{\beta}_I^* = \hat{R}^{-1}\hat{\beta}_I \) where \( \hat{\beta}_I \) is obtained from the logistic regression of \( Y \) on \( (x^*_i, u_i) \) in the validation study, \( R \) can be estimated as \( R = \hat{\Sigma}_x \hat{\Sigma}_{x_i,u}^{-1} \hat{\Sigma}_{x_i,u} = \sum_{i=n_1+1}^{n}[ (x_i^*, u_i)' - \hat{\mu} ][ (x_i^*, u_i)' - \hat{\mu} ]/(n_2 - 1) \)

for the first \( (p+q) \times (p+q) \) elements and zero otherwise, \( \hat{\mu} = \sum_{i=n_1+1}^{n} (x_i^*, u_i)/n_2, \hat{\mu} = E[(x,u)'] \),

\[
\hat{\Sigma}_x = \hat{\Sigma}_{x_i,u} - \hat{\Sigma}/n_R, \text{ and } \hat{\Sigma} = \sum_{i=n_1+1}^{n_2} \sum_{j=1}^{n_R} (x_{ij}^*-x_i)^*(x_{ij}^*-x_i)^*/[n_2(n_R-1)] \text{ for the first } p \times p \text{ elements and zero otherwise. By the delta method, } V_{L,r,s}^* = (\hat{R}^{-1}\text{Var}(\hat{\beta}_I)\hat{R}^{-1})_{r,s} + \hat{\beta}_I^*V_{\hat{\beta}_I}(A)\hat{\beta}_I^* n_R, r,s = 1,...,p+q,
\]

where \( V_{L,r,s}^* \) is the element in the \( r^{th} \) row and \( s^{th} \) column of \( \text{Var}(\hat{\beta}_I) \), \( \text{Var}(\hat{\beta}_I) \) is obtained from
the validation study logistic regression of $Y$ on $x^*$, $A = \hat{\Sigma}_{x_i^*u} - \hat{\Sigma}_{x_i^*u}^{-1}$, and $\text{Var}(A)$ is given by equation (A5) of Rosner et al.\textsuperscript{9}. Assuming (5) is exact, by generalized least squares theory\textsuperscript{29}, $\hat{\beta}_{RC,i}^*$ is unbiased for $\beta$ and minimum variance among all unbiased linear combinations of $\hat{\theta}_{RC}$ and $\hat{\theta}_I^*$, as long as the inverse variance weights are accurately estimated. It is shown in Appendix 2 that unbiased sampling into the validation study will only occur when sampling is independent of $(Y, x_{ij}, ..., x_{i2}, X_i)$ conditional upon $u_i$, for all $i$, $i=1,...,n_1+n_2$. An unbiased sampling mechanism is one in which the usual estimates are consistent without adjustment for sampling.

3 ILLUSTRATIVE EXAMPLES

3.1 The ACE Study of the acute health effects of occupational exposure to anti-neoplastics among pharmacists

Valanis et al. described a cross-sectional study of acute health effects to occupational chemotherapeutics exposures in hospital pharmacists\textsuperscript{28}. Average weekly chemotherapeutics exposure ($X$) was self-reported on questionnaire. In a sub-sample of 56 pharmacists, on-site drug mixing diaries were kept for 1-2 weeks, and the average of daily mixing activities was used as the reference instrument in this validation study ($x$). The correlation between these two methods of exposure assessment was 0.70. Among these 56, 38 had recorded drug mixing activities on two or more days, permitting the assessment of within-person variability in the reference instrument. A research objective was estimation and inference about the prevalence odds ratio for acute health effects related to chemotherapeutics exposure. Here, we will focus on fever prevalence in relation to exposure among 675 pharmacists. There were 110 cases of fever, 5 of
which occurred in the validation study.

The uncorrected analysis of these data were published by Valanis et al.\textsuperscript{28}, and Spiegelman and Casella bias-corrected these results using $\hat{\beta}_{RC}$ and maximum likelihood methods\textsuperscript{29}. Table 1 compares standard methods to the enhancements to regression calibration discussed in this paper. The results from the uncorrected regression (UC) of $Y$ on $X$ and two other covariates taken to be perfectly measured (a binary indicator for night or rotating shift work and age in years) use the main and validation study data together in a ordinary logistic regression model. The multivariate version of $\hat{\beta}_{RC}$ and its associated inferential quantities as given by the multivariate generalization\textsuperscript{8} of equation (2) are denoted by RC. When the robust covariance matrix estimate for the uncorrected logistic regression coefficient $\hat{\beta}$ is used (equation (6)), $\hat{\beta}_{RC,R}$ is obtained (RC,R). Finally, the results for the extension to regression calibration which uses the internal validation study data are given, both with and without the robust variance (RC,I and RC,I,R). The column labeled I refers to the standard logistic regression analysis of $Y$ on $X$ in the 38 validation study subjects with replicate daily work diary data, including 5 cases. The point estimate ($\hat{\beta}$), its estimated standard error ($SE(\hat{\beta})$), the odds ratio and its 95% confidence interval for a 34 dose per week increase in anti-neoplastic mixing activities corresponding to a change from the observed 10th to 90th percentiles of $x$ (OR, 95% CI), and the $p$-value from the Wald test for $H_0: \beta=0$ are given. The bias-corrected point estimates were all three-fold larger than their uncorrected counterpart. Note that the robust variance estimate is approximately 10% smaller than the usual estimate. We are convinced that this is optimistic because the robust variance estimator has been found to be more variable than the standard counterpart\textsuperscript{30}. Because the $\hat{\beta}_{RC,I}$ was
larger than $\hat{\beta}_{RC}$ and the variance estimate is a function of the point estimate, it was not possible to precisely evaluate if any improvement in precision was gained by the RC,I estimator. Because of the small number of cases in the validation study, adjustment for bias in $\hat{\beta}_I$ to obtain $\hat{\beta}_{RC,I}$ had virtually no impact on the results for these data.

3.1 The Nurses’ Health Study of the relationship between dietary Vitamin A intake and breast cancer incidence rates

Hunter et al. described a prospective study of the relationship between breast cancer incidence and average daily vitamin A intake with supplements at baseline among 89,502 women aged 34 to 59 years who were followed for 8 years beginning in 1980. After updating the data to match the analysis given in, 1466 cases occurred during the study period. The logistic regression model used for the analysis adjusted for the effects of total energy intake and alcohol intake, both of which are also measured with error, and for age. Data on the reproductive risk factors for breast cancer, body mass index and a history of benign breast disease were also available. Nutrient intake values were calculated from a 61-item food frequency questionnaire data, which was validated in a sub-sample of 173 women with four one-week weighed diet records, using the average of the days recorded. Among these 173, 168 had 26 or more records of their daily diet, permitting evaluation of within-person variability. Three cases occurred in the validation study during the 8 year follow-up period.

The logistic regression model $\text{logit}(Y_i) = \beta_0 + \beta_1X_i + \beta_2u_i$ was fit to the data, where $Y_i$ is the probability that participant $i$ has received a diagnosis of breast cancer between the time of the 1980 questionnaire return and January 1,1989, $X_i$ are the covariates measured with error,
(alcohol, vitamin A and total energy intake), and $u_i$ other covariates, taken to be perfectly measured (age). Due to supplementation, total vitamin A intake ranged between 276 and 277,183 IU/day in these data, and the data fit the model best when Vitamin A was represented on the log scale. Log$_{10}$ vitamin A and total energy were measured with considerable error; the correlations of values obtained from the food frequency questionnaire with values obtained from the 28 day diet record were 0.42 and 0.36 for log vitamin A and energy, respectively. Including energy in the model in this manner is one of several methods proposed for energy adjustment in nutritional epidemiology$^{33}$, and has the advantage that bias due to measurement error in total energy intake is directly corrected. Alcohol was measured quite well, and the correlation between values obtained from the food frequency questionnaire and the diet record was 0.85. Although other models have been proposed for these and similar data$^{27}$, it is assumed that the measurement error models considered in this paper validly apply to these data$^{34}$. Other measured risk factors included age at menarche, menopausal status, age at first live birth, history of benign breast disease, family history of breast cancer, body mass index, and parity. Because the uncorrected estimates adjusted only for age were essentially the same as the uncorrected estimates adjusted for all measured breast cancer risk factors, we did not adjust for these other risk factors in the analysis that follows. Of course, if these variables were measured with considerable error they could falsely appear to be neither breast cancer risk factors nor confounders of estimated dietary effects on breast cancer risk. The available data indicate that these factors are measured with little error, relative to the amount of error in dietary variables$^{35,36,37,38}$.

The uncorrected analysis of this was first published by Hunter et al.$^{31}$ and the standard regression calibration analysis was given by Spiegelman et al.$^{11}$. We now present further regression calibration analyses using the new methods described above. Table 2 compares standard methods to enhancements using the robust variance for the primary regression model.
coefficients and to enhancements which use the information on $\beta$ from the internal validation study, when available. Similar to what we saw in the previous example, neither enhancement made any material difference to the point and interval estimates of any of the three variables measured with error. Because the internal validation study has only three cases, very little if any gain in efficiency can be obtained from this information. Similarly, adjustment for bias in $\hat{\beta}_j$ to obtain $\hat{\beta}^{*}_{RC,j}$ had no impact on the results for these data.

4 SIMULATION STUDY

4.1 Design of the simulation study

Spiegelman et al. 39 presented a simulation study of the bias, mean-squared error, size, power and coverage probability of the maximum likelihood estimators and related quantities for the parameters of a multiple logistic regression model with covariate misclassification and measurement error. Likelihood-based quantities were compared to the regression calibration method. Although, of course, the MLE and associated quantities are asymptotically optimal, this simulation study showed that in the region of the multi-dimensional parameter space investigated, for small validation study sizes such as typically encountered in practice, standard regression calibration estimates of the multiple logistic regression model parameters had bias equivalent to, or somewhat less than, the MLE, especially near the null. However, in nearly all cases examined, inference using the standard RC approach was inferior to likelihood-based inference, with a conservative nominal size and poor power. These results motivated our further investigation into possibly improved inference using the robust variance for $\hat{\beta}_{RC}$ and to develop $\hat{\beta}_{RC,j}$. The goal of this simulation study was to investigate the extent to which these modifications improved the
performance of the standard regression calibration approach, as compared to likelihood-based methods.

We investigated the extent to which extensions to \( \hat{\beta}_{RC} \) proposed in this paper improve the properties of estimation and inference based upon \( \hat{\beta}_{RC} \), as a function of the main study sample size \((n_1)\), validation study sample size \((n_2)\), the strength of the association of outcome with a continuous variable measured with error, as measured by the log odds ratio, \( \beta_1 \), and the marginal frequency of outcome, as represented by \( \beta_0 \). The design of this simulation study is identical to that given in further detail by Spiegelman et al. 39.

We designed this simulation study to closely follow the data and models given by the data presented in Section 3.2. This gave us a practical basis for choosing values for the twenty model parameters, which define the relative risks of the model covariates, two of which are binary \((W)\) and one of which is continuous \((X)\), the extent of their measurement error and misclassification, and other features. The design matrix, i.e. \((X_i,W_i), i=1,\ldots,n_1\), and \((X_i,W_i), i=n_1+1,\ldots,n_1+n_2\), used here was directly taken from Nurses’ Health Study 40, but also resembles that found in other large, prospective cohort studies of diet and cancer in women currently ongoing 41. Following the Nurses’ Health Study, in which \( x_i^* \) is an average over 26 days or more of weighed food records and, although not a perfect measure of \( x_i \), these averages are close (the reliability coefficients corresponding to the 26-day average were 0.94 and 0.90 for total energy intake and \( \log_{10}(\text{total vitamin A}) \), respectively), simulations for \( \hat{\beta}_{RC,i} \) were not conducted. Each design point was replicated 1825 times. We investigated validation study sizes similar to the Nurses’ Health Study and other prospective studies of diet and cancer \((n_2=173)\) 32 41. Since the performance of the estimators considered was often sub-optimal with validation studies of this standard size, we
sought to determine the effect of doubling and increasing fivefold the validation study size ($n_2$).

To investigate the gain obtained by adding $n_1=8953$ main study subjects to a validation study alone, designs with $n_1=0$ were studied. The design $(n_1=0, n_2=8953)$ represented the scenario when no data are measured with error or mis-classified. In practice, this scenario is unlikely to occur, but it is a useful reference against which losses in power and other adverse effects of measurement error can be gauged. Because conclusions were similar, we report results only on a subset of the design points investigated in this simulation study.

4.2 Results of the simulation study

We present results of the simulation study for $\beta_1$, with $\beta_{21}=0.4055$ and $\beta_0=-2.633$. Patterns described below were similar for the other scenarios considered, and were similar to results observed for $\beta_{21}$ and $\beta_{22}$. Table 3 gives the bias and mean-squared error (MSE) of the estimator of $\beta_1$, the continuous covariate measured with considerable error. In the scenarios considered in our simulation study, $\hat{\beta}_{RC,J}$ had equivalent or less bias than $\hat{\beta}_{RC}$, and usually substantially so. Near the null and at the standard ($n_2=173$) validation study size, $\hat{\beta}_{RC,J}$ had somewhat greater bias and similar mean-squared error than the MLE (ML) for the estimation of $\beta_1$. Table 4 gives the power and size of Wald-type hypothesis tests based upon $\hat{\beta}_{RC,J}$ and $\hat{\beta}_{RC}$ with and without the robust variance. It is clear that the robust variance offered no improvement to $\hat{\beta}_{RC,J}$ with the original variance for $\hat{\beta}_{RC}$ in this setting. Hypothesis tests based upon $\hat{\beta}_{RC,J}$ were still conservative under the null in some cases, but in others, the nominal size was now adequate, offering an improvement to $\hat{\beta}_{RC}$. Power was dramatically improved -- in nearly all cases considered, the
power of $\hat{\beta}_{RC,J}$ was approximately equivalent to that of the ML estimator. Although the uncorrected test based upon $\hat{\beta}$ was the most powerful among all procedures considered in this simulation study, it had incorrect nominal size. Table 5 presents the empirical coverage probability and confidence interval width for Wald-type intervals based upon these estimators. The coverage probability of confidence intervals based upon $\hat{\beta}_{RC,J}$ had the correct size in many instances when the coverage probability of $\hat{\beta}_{RC}$ was incorrect. The coverage probability for $\hat{\beta}_{RC,J}$ was conservative at the standard validation study size. Doubling of the validation study size eliminated this problem. Confidence intervals based upon $\hat{\beta}_{RC,J}$ were narrower than those based upon $\hat{\beta}_{RC}$ in all cases considered, and were typically as efficient as those based upon the ML.

Clearly, $\hat{\beta}_{RC,J}$ is an improvement over $\hat{\beta}_{RC}$ whenever internal validation data is available, even if the validation study is the standard size. For larger validation study sizes, $\hat{\beta}_{RC,J}$ was nearly Fisher efficient and performed as well as the MLE. There was no evidence that use of the robust variance improved the behavior of any of the inferential quantities, confirming our conjecture. Results from the simulation study of the estimators of $\beta_{21}$ were similar to those presented (data not shown). Although the asymptotics for $\hat{\beta}_{RC,J}^*$ were not verified by simulation, we suspect that the results of such a simulation study might be similarly good.

5 Discussion and Conclusion
By example, intuition and through an extensive simulation study, we have shown in this paper that when the disease is rare, the robust variance estimator is unlikely to improve the validity of inference with regression calibration in logistic regression. With the disease is rare, model (5) is nearly correctly specified and in this case, the robust variance should perform somewhat worse in small samples due to its greater instability\(^47\). Using the information about \(\beta\) contained in internal validation studies makes little improvement to the efficiency of the regression calibration estimator with typical validation study sizes, where \(\hat{\beta}_I\) is very variable. When \(\hat{\beta}_I\) is very variable, \(\hat{\beta}_{RC,I}\) could be less efficient than \(\hat{\beta}_{RC}\), although this did not happen in our two examples to any appreciable extent. With larger validation study sizes, our simulation study showed that in contrast to the standard regression calibration estimator, \(\hat{\beta}_{RC}\), the regression calibration estimator proposed in this paper, \(\hat{\beta}_{RC,I}\), was nearly unbiased and Fisher efficient, at least in the region of the parameter space that was investigated.

Application of the methods developed in Section 2.2 did not alter results in either study to which it was applied. The validation study sizes in both of these studies were small and few cases had occurred in either. In the Nurses’ Health Study validation study, a large number of replicates for each individual were available. Hence, the adjustment for bias due to random within-person variation around an intra-individual mean over 26 days is unlikely to have appreciable impact. In other settings, this development can be expected to have much greater impact when the validation study size is large, the event of interest more common, and the number of replicates within individual small. On the other hand, standard regression calibration methods such as given by Rosner et al.\(^8\) will often be adequate.
Regression calibration via $\hat{\beta}_{RC}$ of main study/validation study data is a useful applied tool for measurement error correction. It is increasingly being adopted by epidemiologists and other biomedical scientists in routine study design and analysis, particularly, thus far, in the field of nutritional epidemiology. It is appealing for use in applications because investigators may proceed with their usual analytic methods, adjusting their final point and interval estimates by a non-iterative procedure with user-friendly SAS macros which are publicly available*.

Regression calibration was originally proposed under the assumption that a "gold standard" method of exposure assessment is available in the validation study. As discussed by Wacholder et al.20, this assumption is rarely if ever met -- rather, "alloyed" gold standards are available. Fortunately, if the errors in the alloyed gold standard are uncorrelated with perfectly measured model covariates and the errors in the usual method of exposure assessment, the regression calibration estimator remains approximately consistent, although its variance will be larger than if a true gold standard were available42. Methods have been developed in the present paper to use the data on the exposure-disease association from the validation study, when an imperfect, unbiased reference instrument is available. Under certain assumptions which may be met in many instances, methods have been developed to empirically verify the assumption of uncorrelated errors and to modify regression calibration when that assumption is not met 42. It should be noted that in nutritional epidemiology, such as in the Nurses’ Health Study data presented in this paper, there is conflicting data as to the structure of errors in dietary assessment instruments (see 27 43 and 44 for further discussion of this point).

Regression calibration via $\hat{\beta}_{RC}$ offers flexibility in the measurement error models which

* Request via electronic mail to Dr. Spiegelman at stdls@channing.harvard.edu
Regression calibration falls somewhere in between the fully parametric and semi-
parametric approaches, because it specifies only the first two moments of the distribution of \( x \) given \( X \). In our simulations, the regression calibration approach was nearly fully efficient at validation study sizes of around 350 observations -- of course, the simulation design satisfied the assumptions of regression calibration that the mean of \( x \) given \( X \) was linear and the variance of \( x \) given \( X \) was constant. In addition, the outcome was relatively rare and the distribution of \( x \) given \( X \) was Gaussian. As the outcome becomes more common, and/or as the distribution of \( x \) given \( X \) departs more substantially from the linear mean, constant variance assumptions, it is possible that the fully parametric and semi-parametric approaches may do better than the version of regression calibration considered in this paper\(^8\,9\,11\), which could become badly biased. This point remains to be investigated. Alternatively, more complex mean and variance functions could be specified, and the regression calibration approach detailed by Carroll et al.\(^4\) could be applied. If the departure from the assumptions involves introduction of heterogeneity to \( \text{Var}(x \mid X) \), a further extension to Rosner et al.’s regression calibration estimator is available\(^{49}\).

Just as one cannot control for confounding without collecting data on confounders, one cannot correct for bias due to measurement error without conducting a validation sub-study. If the reference instrument is imperfect but unbiased with errors uncorrelated with those of the usual exposure measure, a reliability sub-study is needed as well. Since these studies are expensive to conduct, careful sample size calculations are needed. Guidance on some aspects of efficient study design in this setting is available\(^5\,6\,7\) and many epidemiologists already routinely validate and assess the error in their measurements, e.g.\(^{32}\,50\,51\,52\,53\,54\). Further research is needed, particularly when an imperfect, unbiased reference instrument is expected. We hope that this paper will serve to encourage more extensive use of the data obtained from main study/validation study designs, in producing less biased and more precise estimates of health effects, with confidence intervals that correctly reflect the true uncertainty in the data.
Table 1. Results from the ACE Study\(^1\)

<table>
<thead>
<tr>
<th></th>
<th>(\hat{\beta} )</th>
<th>(SE(\hat{\beta}))</th>
<th>OR(^2)</th>
<th>95% CI(^2)</th>
<th>(p)-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\hat{\beta}_{UC} )(^3)</td>
<td>0.0164</td>
<td>0.0064</td>
<td>1.083</td>
<td>1.02-1.15</td>
<td>0.010</td>
</tr>
<tr>
<td>(\hat{\beta}_{RC} )</td>
<td>0.0506</td>
<td>0.0290</td>
<td>1.279</td>
<td>0.97-1.64</td>
<td>0.081</td>
</tr>
<tr>
<td>(\hat{\beta}_{RC,R} )</td>
<td>same as RC</td>
<td>0.0260</td>
<td>same as RC</td>
<td>1.00-1.64</td>
<td>0.051</td>
</tr>
<tr>
<td>(\hat{\beta}_I )</td>
<td>0.0131</td>
<td>0.1495</td>
<td>1.07</td>
<td>0.26-4.42</td>
<td>0.894</td>
</tr>
<tr>
<td>(\hat{\beta}_I^* )</td>
<td>0.0260</td>
<td>0.2968</td>
<td>1.13</td>
<td>0.07-19.1</td>
<td>0.930</td>
</tr>
<tr>
<td>(\hat{\beta}_{RC,I} )</td>
<td>0.0494</td>
<td>0.0284</td>
<td>1.27</td>
<td>0.97-1.67</td>
<td>0.082</td>
</tr>
<tr>
<td>(\hat{\beta}_{RC,I}^* )</td>
<td>0.0518</td>
<td>0.0285</td>
<td>1.29</td>
<td>0.98-1.69</td>
<td>0.069</td>
</tr>
<tr>
<td>(\hat{\beta}_{RC,I,R} )</td>
<td>0.0498</td>
<td>0.0256</td>
<td>1.27</td>
<td>1.00-1.62</td>
<td>0.052</td>
</tr>
<tr>
<td>(\hat{\beta}_{RC,I,R}^* )</td>
<td>0.0519</td>
<td>0.0256</td>
<td>1.29</td>
<td>1.01-1.64</td>
<td>0.042</td>
</tr>
</tbody>
</table>

\(^1\) All results adjusted for age in years and shift (night or rotating vs. otherwise)

\(^2\) OR, 95% CI correspond to odds ratio and 95% confidence interval for a 34 dose/week increase in mixing activity

\(\hat{\beta}_{UC} = \text{uncorrected}\)
\(\hat{\beta}_{RC} = \text{regression calibration}\)
\(\hat{\beta}_{RC,R} = \text{regression calibration with robust variance}\)
\(\hat{\beta}_I = \text{internal validation alone}\)
\(\hat{\beta}_I^* = \text{internal validation with correction for random within-person variability}\)
\(\hat{\beta}_{RC,I} = \text{regression calibration with internal validation}\)
\(\hat{\beta}_{RC,I}^* = \text{same as }\hat{\beta}_{RC,I} \text{ but using }\hat{\beta}_I^* \text{ instead of }\hat{\beta}_I\)
\(\hat{\beta}_{RC,I,R} = \text{same as }\hat{\beta}_{RC,I} \text{ but with robust variance for }\hat{\beta}_{UC}\)
\(\hat{\beta}_{RC,I,R}^* = \text{same as }\hat{\beta}_{RC,I}^* \text{ but with robust variance for }\hat{\beta}_{UC}\)
Table 2. Results from the Nurses’ Health Study$^1$

<table>
<thead>
<tr>
<th>variable</th>
<th>$\hat{\beta}_{UC}^2$</th>
<th>$\hat{\beta}_{RC}$</th>
<th>$\hat{\beta}_{RC,R}$</th>
<th>$\hat{\beta}_{I}$</th>
<th>$\hat{\beta}_{I}^*$</th>
<th>$\hat{\beta}_{RC,I}$</th>
<th>$\hat{\beta}_{RC,I}^*$</th>
<th>$\hat{\beta}_{RC,I,R}$</th>
<th>$\hat{\beta}_{RC,I,R}^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\beta}$</td>
<td>alcohol</td>
<td>0.00787</td>
<td>-0.00003</td>
<td>-0.2398</td>
<td>same as RC</td>
<td>-0.1853</td>
<td>-0.1837</td>
<td>0.01339</td>
<td>0.01382</td>
</tr>
<tr>
<td></td>
<td>energy</td>
<td>0.000053</td>
<td>-0.00008</td>
<td>-0.71963</td>
<td>RC</td>
<td>-0.00193</td>
<td>-0.00203</td>
<td>-0.00013</td>
<td>-0.00013</td>
</tr>
<tr>
<td></td>
<td>log$_{10}$(Vitamin A)</td>
<td>0.1042</td>
<td>0.29326</td>
<td>0.30188</td>
<td>5.48</td>
<td>0.00478</td>
<td>0.00524</td>
<td>0.29263</td>
<td>0.29137</td>
</tr>
<tr>
<td>$SE(\hat{\beta})$</td>
<td>alcohol</td>
<td>0.00204</td>
<td>0.00034</td>
<td>0.1042</td>
<td>0.2109</td>
<td>0.2135</td>
<td>0.00374</td>
<td>0.00374</td>
<td>0.00365</td>
</tr>
<tr>
<td></td>
<td>energy</td>
<td>0.000053</td>
<td>0.00021</td>
<td>0.29326</td>
<td>0.00020</td>
<td>0.00478</td>
<td>0.00021</td>
<td>0.29263</td>
<td>0.29137</td>
</tr>
<tr>
<td></td>
<td>log$_{10}$(Vitamin A)</td>
<td>0.1042</td>
<td>0.29326</td>
<td>0.30188</td>
<td>5.48</td>
<td>0.00478</td>
<td>0.00524</td>
<td>0.29263</td>
<td>0.29137</td>
</tr>
<tr>
<td>OR$^3$</td>
<td>alcohol</td>
<td>1.10</td>
<td>1.17</td>
<td>same as RC</td>
<td>0.38</td>
<td>0.11</td>
<td>1.17</td>
<td>1.18</td>
<td>1.18</td>
</tr>
<tr>
<td></td>
<td>energy</td>
<td>0.98</td>
<td>0.94</td>
<td>RC</td>
<td>0.69</td>
<td>0.20</td>
<td>0.92</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>log$_{10}$(Vitamin A)</td>
<td>0.93</td>
<td>0.81</td>
<td>0.23</td>
<td>0.79</td>
<td>0.78</td>
<td>0.79</td>
<td>0.78</td>
<td>0.78</td>
</tr>
<tr>
<td>95% CI$^3$</td>
<td>alcohol</td>
<td>1.05-1.15</td>
<td>1.08-1.28</td>
<td>1.08-1.28</td>
<td>1.08-1.28</td>
<td>1.08-1.28</td>
<td>1.08-1.28</td>
<td>1.08-1.28</td>
<td>1.08-1.28</td>
</tr>
<tr>
<td></td>
<td>energy</td>
<td>0.90-1.06</td>
<td>0.67-1.31</td>
<td>0.68-1.29</td>
<td>0.66-1.29</td>
<td>0.65-1.26</td>
<td>0.67-1.27</td>
<td>0.66-1.24</td>
<td>0.65-1.24</td>
</tr>
<tr>
<td></td>
<td>log$_{10}$(Vitamin A)</td>
<td>0.87-0.99</td>
<td>0.68-0.96</td>
<td>0.68-0.96</td>
<td>0.68-0.96</td>
<td>0.67-0.94</td>
<td>0.67-0.94</td>
<td>0.66-0.95</td>
<td>0.65-0.93</td>
</tr>
<tr>
<td>p-value</td>
<td>alcohol</td>
<td>0.0001</td>
<td>0.0004</td>
<td>0.0003</td>
<td>0.39</td>
<td>0.0003</td>
<td>0.0002</td>
<td>0.0002</td>
<td>0.0001</td>
</tr>
<tr>
<td></td>
<td>energy</td>
<td>0.5983</td>
<td>0.71</td>
<td>0.70</td>
<td>0.70</td>
<td>0.64</td>
<td>0.54</td>
<td>0.63</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>log$_{10}$(Vitamin A)</td>
<td>0.0214</td>
<td>0.0141</td>
<td>0.0171</td>
<td>0.42</td>
<td>0.44</td>
<td>0.0087</td>
<td>0.0044</td>
<td>0.0107</td>
</tr>
</tbody>
</table>

$^1$ Results adjusted for age as a continuous covariate

$^2$ $\hat{\beta}_{UC}$: uncorrected

$\hat{\beta}_{RC}$: regression calibration

$\hat{\beta}_{RC,R}$: regression calibration with robust variance

$\hat{\beta}_{I}$: internal validation alone

$\hat{\beta}_{I}^*$: internal validation with correction for random within-person variability

$\hat{\beta}_{RC,I}$: regression calibration with internal validation

$\hat{\beta}_{RC,I}^*$: same as $\hat{\beta}_{RC,I}$ but using $\hat{\beta}_{I}^*$ instead of $\hat{\beta}_{I}$

$\hat{\beta}_{RC,I,R}$: same as $\hat{\beta}_{RC,I}$ but with robust variance for $\hat{\beta}_{UC}$

$\hat{\beta}_{RC,I,R}^*$: same as $\hat{\beta}_{RC,I,R}$ but with robust variance for $\hat{\beta}_{UC}$

$^3$ Odds ratio (OR) and 95% confidence interval (95% CI) corresponding to a 12g/day increase in alcohol, a 800 kcal/day increase in energy, and a 0.3 log IU/day increase in Vitamin A
Table 3: Bias and mean-squared error for $\beta_1$  

<table>
<thead>
<tr>
<th>$n_2$</th>
<th>0</th>
<th>173</th>
<th>173'2</th>
<th>8953</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_\text{UC}$ = Uncorrected, $\beta_\text{RC}$ = Regression calibration estimator, $\beta_{\text{RC}, R}$ = Regression calibration with robust variance, $\beta_{\text{RC}, I}$ = Efficient regression calibration estimator, $\beta_{\text{RC}, I, R}$ = Efficient regression calibration estimator with robust variance, ML = Maximum likelihood estimator, with main study/internal validation study design, TR = Maximum likelihood estimator, without measurement error/misclassification, Absolute Bias, MSE, Relative Bias, MSE.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. $\beta_\text{UC}$ = Uncorrected, $\beta_\text{RC}$ = Regression calibration estimator
2. $\beta_{\text{RC}, R}$ = Regression calibration with robust variance
3. $\beta_{\text{RC}, I}$ = Efficient regression calibration estimator
4. $\beta_{\text{RC}, I, R}$ = Efficient regression calibration estimator with robust variance
5. ML = Maximum likelihood estimator, with main study/internal validation study design
6. TR = Maximum likelihood estimator, without measurement error/misclassification
7. Absolute Bias, MSE
8. Relative Bias, MSE

| $n_1$ = 0 TR | 0.012, 1.50 | 0.022, 0.67 | -0.000, 0.02 |
| $n_1 = 8953$ $\beta_\text{RC}$ | 0.063, 0.52 | 0.057, 0.33 |
| $\beta_{\text{RC}, R}$ | 0.063, 0.52 | 0.057, 0.33 |
| $\beta_{\text{RC}, I}$ | 0.073, 0.28 | 0.056, 0.20 |
| $\beta_{\text{RC}, I, R}$ | 0.073, 0.28 | 0.056, 0.20 |
| ML | -0.013, 0.49 | -0.009, 0.23 |

| $n_1 = 8953$ $\beta_\text{RC}$ | 0.024, 0.58 | 0.022, 0.27 | -0.001, 0.01 |
| $\beta_{\text{RC}, R}$ | 0.145, 0.88 | 0.127, 0.25 |
| $\beta_{\text{RC}, I}$ | 0.145, 0.88 | 0.127, 0.25 |
| $\beta_{\text{RC}, I, R}$ | 0.041, 0.19 | 0.053, 0.11 |
| ML | 0.048, 0.26 | 0.024, 0.14 |

file:/udd/strol/donna-misc/newtables/sic3.tex
Table 4: Size and power of $H_0: \beta_1 = 0$

<table>
<thead>
<tr>
<th>$n_2$</th>
<th>0</th>
<th>173</th>
<th>173*2</th>
<th>8953</th>
</tr>
</thead>
</table>

| $\beta_1 = 0.0$ |  |  |
| n1=0 | TR | 0.05² |
| n1=8953 | $\hat{\beta}_{UC}$ | 0.13 |
| | $\hat{\beta}_{RC}$ | 0.02 | 0.04 |
| | $\hat{\beta}_{RC,R}$ | 0.02 | 0.04 |
| | $\hat{\beta}_{RC,I}$ | 0.03 | 0.05 |
| | $\hat{\beta}_{RC,I,R}$ | 0.02 | 0.05 |
| | ML | 0.04 | 0.05 |

| $\beta_1 = 0.8109$ |  |  |
| n1=0 | TR | 1.00⁹ |
| n1=8953 | $\hat{\beta}_{UC}$ | 1.00 |
| | $\hat{\beta}_{RC}$ | 0.33 | 0.54 |
| | $\hat{\beta}_{RC,R}$ | 0.33 | 0.54 |
| | $\hat{\beta}_{RC,I}$ | 0.48 | 0.73 |
| | $\hat{\beta}_{RC,I,R}$ | 0.48 | 0.73 |
| | ML | 0.45 | 0.66 |

1. TR = Maximum likelihood, no measurement error/misclassification,
   $\hat{\beta}_{UC}$ = Uncorrected,
   $\hat{\beta}_{RC}$ = Wald test based on regression calibration estimator
   $\hat{\beta}_{RC,R}$ = Wald test based on regression calibration estimator with robust variance
   $\hat{\beta}_{RC,I}$ = Wald test based on efficient regression calibration estimator
   $\hat{\beta}_{RC,I,R}$ = Wald test based on efficient regression calibration estimator with robust variance
   ML = Wald test using observed Fisher information variance and main study/internal validation study design

2. Size
3. Shaded data points in the design/parameter space which fall outside the 95% confidence limits for the expected size (0.05)
4. Power
Table 5: Coverage Probability and Confidence Interval Width $\beta_1$  

<table>
<thead>
<tr>
<th>$n_2$</th>
<th>0</th>
<th>173</th>
<th>173*2</th>
<th>8953</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta = 0.0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n_1 = 0$</td>
<td>TR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n_1 = 8953$</td>
<td>$\hat{\beta}_{UC}$</td>
<td>0.87, 1.44</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{\beta}_{RC}$</td>
<td>0.98, 0.21</td>
<td>0.96, 7.56</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{\beta}_{RC,R}$</td>
<td>0.98, 0.21</td>
<td>0.96, 7.56</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{\beta}_{RC,I}$</td>
<td>0.97, 7.28</td>
<td>0.95, 5.39</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{\beta}_{RC,I,R}$</td>
<td>0.98, 7.29</td>
<td>0.95, 5.44</td>
<td></td>
</tr>
<tr>
<td>ML</td>
<td></td>
<td></td>
<td>0.96, 8.65</td>
<td>0.95, 5.87</td>
</tr>
<tr>
<td>$\beta = 0.8109$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n_1 = 0$</td>
<td>TR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n_1 = 8953$</td>
<td>$\hat{\beta}_{UC}$</td>
<td>0.00, 1.26</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{\beta}_{RC}$</td>
<td>0.99, 0.63</td>
<td>0.97, 5.73</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{\beta}_{RC,R}$</td>
<td>0.99, 0.63</td>
<td>0.97, 5.73</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{\beta}_{RC,I}$</td>
<td>0.98, 5.88</td>
<td>0.96, 3.70</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{\beta}_{RC,I,R}$</td>
<td>0.98, 5.88</td>
<td>0.96, 3.70</td>
<td></td>
</tr>
<tr>
<td>ML</td>
<td></td>
<td></td>
<td>0.96, 6.36</td>
<td>0.95, 3.98</td>
</tr>
</tbody>
</table>

1. TR = Maximum likelihood, no measurement error/misclassification,  
| $\hat{\beta}_{UC}$ = Uncorrected,  
| $\hat{\beta}_{RC}$ = Wald CI (confidence interval) based on regression calibration estimator,  
| $\hat{\beta}_{RC,R}$ = Wald CI based on regression calibration estimator with robust variance,  
| $\hat{\beta}_{RC,I}$ = Wald CI based on efficient regression calibration estimator,  
| $\hat{\beta}_{RC,I,R}$ = Wald CI based on efficient regression calibration estimator with robust variance  
ML = Wald CI using observed Fisher information variance,  
main study/internal validation study design

2. Empirical Coverage Probability, exp(Upper Bound/Lower Bound)

3. Shaded data points in the design/parameter space which fall outside the 95% confidence limits for the expected coverage probability (95%).

file://add/strol/donna-misc/newtables/sk5.tex , sk5.ps 

26
Appendix 1. Derivation of the asymptotic covariance of $\hat{\beta}_I$ and $\hat{\beta}_{RC}$

The results derived here are approximate, made under the assumption that model (5) holds exactly. In logistic regression models with a rare outcome, this is very nearly true. For simplicity, we drop the variables $u$ in (5). As in model (3), the regression of $x$ on $X$ has least squares estimate $\hat{\Gamma}$. Since $\hat{\beta}_{RC} = \hat{\Gamma}^{-1} \hat{\beta}$ is a function of the validation data only through $\hat{\Gamma}$, we need to show that $\hat{\Gamma}$ and $\hat{\beta}_I$ are asymptotically uncorrelated as the validation sample size $n_2 \to \infty$.

By asymptotically uncorrelated, we mean that

$$\sqrt{n_2} \begin{bmatrix} \hat{\beta}_I - \beta \\ \text{vec}(\hat{\Gamma} - \Gamma) \end{bmatrix}$$

is asymptotically multivariate normally distributed with mean zero and a covariance matrix which is of block diagonal form, where $\text{vec}(\Gamma)$ is the vectorized form of the matrix $\Gamma$. For all generalized linear models, including logistic regression, linear regression and Cox proportional hazards, as well as for models considered by Carroll and Ruppert, it is well known that, for some function $\psi_1(Y, x; \beta)$, the estimated primary regression slope $\hat{\beta}_I$ has the expansion

$$\sqrt{n_2} (\hat{\beta}_I - \beta) = n_2^{-\frac{1}{2}} \sum_{i=1}^{n_2} \psi_1(Y_i, x_i; \beta) + o_p(1),$$

where $\psi_1(Y_i, x_i; \beta)$ is defined such that

$$E\{\psi_1(Y, x; \beta) | x\} = E\{\psi_1(Y, x; \beta) | (x, X)\} = 0. \quad (A1.1)$$

The first equality in (A1.1) follows from the fact that since $X$ is a surrogate, $Y$ and $X$ are independent given $x$. In the validation study, the estimated regression slope obtained from the linear measurement error model (3) has a similar expansion

$$\sqrt{n_2} \text{vec}(\hat{\Gamma} - \Gamma) = n_2^{-\frac{1}{2}} \sum_{i=1}^{n_2} \psi_2(x_i, X_i; \Gamma) + o_p(1), \quad (A1.2)$$

for a function $\psi_2(x, X; \Gamma)$ defined such that

$$E\{\psi_2(x, X; \Gamma) | X\} = 0.$$

In the present setting, $\psi_1(Y_i, x_i; \beta)$ is the standard score function from the logistic regression model (6) for $Y$ given $x$, and $\psi_2(x, X; \Gamma)$ is the vectorized version of the normal equations for regressing
With these facts, we can now show that \( \hat{\beta}_I \) and \( \hat{\Gamma} \) are asymptotically uncorrelated. By the theory of estimating equations (Carroll et al.), bivariate asymptotic normality follows, with the covariance of this asymptotic distribution given by \( A'BA' \), where

\[
A = \begin{bmatrix}
E \frac{\partial \psi_1(Y,x;\beta)}{\partial \beta} & 0 \\
0 & E \frac{\partial \psi_2(x,X;\Gamma)}{\partial \text{vec}(\Gamma)}
\end{bmatrix}
\]

and

\[
B = E \left[ \begin{pmatrix}
\psi_1(Y,x;\beta) \\
\psi_2(x,X;\Gamma)
\end{pmatrix} \begin{pmatrix}
\psi_1(Y,x;\beta) \\
\psi_2(x,X;\Gamma)
\end{pmatrix}^T \right].
\]

The main result thus follows if we can show that

\[
E(\psi_1(Y,x;\beta) \psi_2^T(x,X;\Gamma)) = 0.
\]

But this is a simple consequence of (A1.1) and (A1.2) since

\[
E(\psi_1(Y,x;\beta) \psi_2^T(x,X;\Gamma)) = E\left[ E(\psi_1(Y,x;\beta) \psi_2^T(x,X;\Gamma) | x,X) \right] = E\left[ E(\psi_1(Y,x;\beta) | x,X) \psi_2^T(x,X;\Gamma) \right] = E(0 \times \psi_2^T(x,X;\Gamma)) = 0.
\]
Appendix 2. Definition of unbiased sampling into the validation study

Let $\Delta_i=1$ if data for subject $i$ is $(Y_i, X_i, u_i)$, $\Delta_i=2$ if data for subject $i$ is $(Y_i, X_i, x_i, u_i)$, and let $\Delta_i=3$ if data for subject $i$ is $(Y_i, X_i, x_i^*, ..., x_i^n_{R_i}, u_i)$. $\hat{\beta}_{RC,J}$ is estimated when all subjects have $\Delta=1$ or $\Delta=2$, and $\hat{\beta}_{RC,J}^*$ is estimated when $\Delta=1$ or $\Delta=3$ for all subjects. Then there are parameters $\Gamma$ and $S$ such that $E(x | X, u; \Gamma) = E(x^* | X, u; \Gamma) = m_1(X, u; \Gamma)$ and $E(x | x^*, u; S) = m_2(x^*, u; S)$. There are three major estimating functions to consider:

1. $\psi_1(Y, x, u; \beta)$ for estimating $\beta$, where
   \[ E\{\psi_1(Y, x, u; \beta) | x, u\} = E\{\psi_1(Y, x, u; \beta) | x, u, x^*\} = 0, \]

2. $\psi_2(x, X, u; \Gamma)$ for estimating $\Gamma$, where
   \[ E\{\psi_2(x, X, u; \Gamma) | X, u\} = 0, \]  
   (1)

3. $\psi_3(x^*, X, u; S)$ for estimating $S$, where
   \[ E\{\psi_3(x^*, X, u; S) | x^*, u\} = 0. \]

In addition, note that the regression calibration approximation states that

\[ E\{\psi_1(Y, m_1(X, u; \Gamma), u; \beta) | X, u\} = E\{\psi_1(Y, m_2(x^*, u; S), u; \beta) | x^*, u\} = 0. \]

Note importantly that, because $E(Y | X, x^*, u) \neq E(Y | x^*, u)$,

\[ E\{\psi_1(Y, m_2(x^*, u; S), u; \beta) | X, x^*, u\} \neq 0. \]  
   (2)

Assuming that selection into the validation study depends only on $(X, U)$, we define

\[ \pi(\delta, X, u) = Pr(\Delta = \delta | X, u) = Pr(\Delta = \delta | x, u, X, x^*, Y). \]  
   (3)

Let $\Omega=1$ if $\hat{\beta}_{J}^*$ is used to estimate $\beta$ when $\Delta=3$ and let $\Omega=0$ if the regression calibration method of Rosner et al.\(^8\) is used to estimate $\beta$ among the validation study subjects when $\Delta=3$. Note that $\Omega$ does not depend upon the data in any way. The estimating equations for the data are thus
For consistency, we must show that these estimating equations are unbiased, i.e. have mean zero under the sampling plan. Using (3), it is easily seen that the latter two equations above are unbiased. For example, from (1),

\[ E[I(\Delta_i=2)\psi_2(x_i,X_i,u_i;\Gamma)] = E[E[I(\Delta_i=2)\psi_2(x_i,X_i,u_i;\Gamma)] | x_i,X_i,u_i] = E[\pi(2,X,u)\psi_2(x_i,X_i,u_i;\Gamma)] = E[\pi(2,X,u)\psi_2(x_i,X_i,u_i;\Gamma)] | X_i,u_i = 0. \]

Showing (5) has mean zero follows in a similar fashion.

From (3) and calculations similar to those given above, it is easily seen that (4) has expectation

\[ \Omega E[\pi(3,X,u)\psi_1(Y_i,m_2(x_i^*,u_i^*;S),u_i;\beta)]. \]

For \( \hat{\beta}_{RC,}\), \( \pi(3,X_i,u_i)=0 \) for all \( i, i=1,...,n_1+n_2 \), so (6) vanishes and (4) has mean zero. For \( \hat{\beta}_{*RC},\Omega=1 \) and (6) has mean zero (within regression calibration approximations) only if \( \pi(3,X,u) = \pi(3,u) \), i.e. only if sampling into the validation study depends, at most, only on \( u \). If instead of \( \hat{\beta}_{RC} \) the regression calibration approach of Rosner et al. 8 is used to estimate \( \beta \) in the validation study, i.e. if \( \Omega=0 \), (4) is unbiased even if sampling into the validation study depends jointly upon \( (X,u) \).
Appendix 3. Derivation of $C\hat{\text{ov}}(\hat{\beta}_{RC}, \hat{\beta}_{M})$

Let $\theta=(\text{vec}(\Gamma), \text{vech}(\Sigma_{x_i|x}), \text{vech}(\Sigma_{x_i,u}), \text{vech}(\Sigma)\beta)'$, where

$E(x_i|x, u) = E(x_i^*|x, u) = (x_i^*, u_i, 1)^T \Gamma$, $\text{Var}(x_i^*|x, u) = \Sigma_{x_i|x}$, $\text{Var}(x_i^*, u) = \Sigma_{x_i,u}$, $\text{Var}(e_{ij}) = \Sigma$.

$E(x_i|U) = E(x_i^*) = \mu$, $E(x_i|U) = E(x_i^*|U) = (x_i^*, u_i, 1)^T \beta_p$, and $E(Y_i|x_i, u_i) = (x_i^*, u_i)\beta_M$, and $\beta_M$ is the convergent value of the main study logistic regression parameter estimates. Then, $\hat{\theta}$ is the solution to $\psi(\theta) = 0$, where $\psi(\theta) = (\psi_1, \psi_2, \psi_3, \psi_4, \psi_5, \psi_6, \psi_7)'$.

$$\psi_1(\theta) = \sum_{i=1}^{n_2} \sum_{j=1}^{n_R} \delta(j=1) \text{vec}((X_i u_i, 1) \otimes (x_i^* - (x_i^*, u_i, 1)^T \Gamma)$$

$$\psi_2(\theta) = \sum_{i=1}^{n_2} \sum_{j=1}^{n_R} \delta(j=1) \text{vech}\{[x_i^* - (X_i u_i, 1)^T \Gamma][x_i^* - (X_i u_i, 1)^T \Gamma]' - \Sigma_{x_i|x}\}$$

$$\psi_3(\theta) = \sum_{i=1}^{n_2} \sum_{j=1}^{n_R} \delta(j=1) \text{vech}\{(x_i^*, u_i)' - \mu)(x_i^*, u_i)' - \mu - \Sigma_{x_i,u}\}$$

$$\psi_4(\theta) = \sum_{i=1}^{n_2} \sum_{j=1}^{n_R} \text{vech}\{(x_i^*, u_i)'(x_i^*, u_i)' - \frac{n_R^{-1}}{n_R} \Sigma\}$$

$$\psi_5(\theta) = \sum_{i=1}^{n_2} \sum_{j=1}^{n_R} \delta(j=1) \ (x_i^*, u_i)' - \mu$$

$$\psi_6(\theta) = \sum_{i=1}^{n_2} \sum_{j=1}^{n_R} \delta(j=1) \ w_n(x_i^*, u_i) [Y_i - (x_i^*, u_i, 1)^T \beta_f]$$

$$\psi_7(\theta) = \sum_{i=n_2+1}^{n_2+n_1} w_M(X_i u_i) [Y_i - (X_i u_i, 1)^T \beta_M]$$

where $\delta(\cdot)$ is an indicator function equal to 1 when the condition inside the parentheses is true and 0 otherwise, $w_n(x_i^*, u_i) = \frac{\partial E(Y_i|x_i^*, u_i, 1)^T \beta_f)}{\partial \beta_f} - (x_i^*, u_i, 1)'$ and

$w_M(X_i u_i) = \frac{\partial E(Y_i|x_i^*, u_i; \beta_M)}{\partial \beta_M} - (X_i u_i, 1)'$. Then, $C\hat{\text{ov}}(\hat{\theta}) = I^{-1} G I^{-1}$, where

31
\[ G = \sum_{i=1}^{n_z} \sum_{j=1}^{n_R} \psi_{ij} \psi_{ij}' \quad \text{and} \quad I = \sum_{i=1}^{n_z} \sum_{j=1}^{n_R} \frac{\partial \psi_{ij}}{\partial \theta} \]  \\
Let \[ g(\theta) = [g_1'(\theta), g_2'(\theta)]' = \left( \begin{bmatrix} \Gamma^{-1} \beta_M \end{bmatrix}', \begin{bmatrix} (L - \frac{\Sigma \Sigma^{-1}_{x_i, u}}{n_R})^{-1} \beta \end{bmatrix}' \right)' = (\beta_{RC}', \beta_I')' , \]  
where, here, \( L \) is the identity matrix of \( \text{dim}[(p+q+1) \times (p+q+1)] \), \( \Sigma \) is augmented to \( \text{dim}[(p+q+1)^2] \) with 0’s outside of the \( p \times p \) upper left block, \( \Sigma_{x_i, u} \) is augmented with an additional column and row of 0’s in the \( (p+q+1) \) position, \( \Gamma \) is augmented to \( \text{dim}[(p+q+1)^2] \) with a \( (q+1)^2 \) identity matrix in the lower diagonal \( (p+1) \) to \( (p+q+1) \) positions and 0’s everywhere else, and \( \beta_M = \Gamma \beta \). Then,

\[ C_0v(\beta_{RC}, \beta_I') = C_0v(g(\theta)) = g_0(\theta) |_{\theta=0} C_0v(\theta) g_0'(\theta) |_{\theta=0} = \begin{bmatrix} Q_1 & V_{RCI}^* \\ Q_2 & \end{bmatrix} \]

since \( C_0v(\hat{\beta}, \hat{\beta}) = \theta \), where \( g_0(\theta) = \begin{bmatrix} \frac{\partial g_1(\theta)}{\partial \theta} \\ \frac{\partial g_2(\theta)}{\partial \theta} \end{bmatrix} \), and \( Q_1 \) and \( Q_2 \) are \( (p+q+1)^2 \) matrices that are not of interest.
References


