Published by Oxford University Press on behalf of the International Epidemiological Association *International Journal of Epidemiology* 2006;**35:**1074–1081

doi:10.1093/ije/dyl097

Multiple-imputation for measurement-error correction

Stephen R Cole,* Haitao Chu and Sander Greenland

Accepted 10 April 2006

Background There are many methods for measurement-error correction. These methods remain rarely used despite the ubiquity of measurement error.

Methods Treating measurement error as a missing-data problem, the authors show how multiple-imputation for measurement-error (MIME) correction can be done using SAS software and evaluate the approach with a simulation experiment.

Results Based on hypothetical data from a planned cohort study of 600 children with chronic kidney disease, the estimated hazard ratio for end-stage renal disease from the complete data was 2.0 [95% confidence limits (95% CL) 1.4, 2.8] and was reduced to 1.5 (95% CL 1.1, 2.1) using a misclassified exposure of low glomerular filtration rate at study entry (sensitivity of 0.9 and specificity of 0.7). The MIME correction hazard ratio was 2.0 (95% CL 1.2, 3.3), the regression calibration (RC) hazard ratio was 2.0 (95% CL 1.1, 3.7), and restriction to a 25% validation substudy yielded a hazard ratio of 2.0 (95% CL 1.0, 3.7). Based on Monte Carlo simulations across eight scenarios, MIME was approximately unbiased, had approximately correct coverage, and was sometimes more powerful than misclassified or RC analyses. Using root mean squared error as a criterion, the MIME bias correction is sometimes outweighed by added imprecision.

Conclusion The choice between MIME and RC depends on performance, ease, and objectives. The usefulness of MIME correction in specific applications will depend upon the sample size or the proportion validated. MIME correction may be valuable in interpreting imperfectly measured epidemiological data.

Keywords Bias, measurement error, misclassification, missing data, multiple-imputation

Mismeasurement of exposure, disease, or covariates is common in epidemiological research. There are many methods for measurement-error correction, all of which use information that maps observed measurements to true values. Examples of these methods include regression calibration (RC) and moment reconstruction,[1–3] in which the mapping is based on a validation study; and sensitivity analysis,[4] in which the mapping may be based on prior information or speculation. Despite the ubiquity of measurement error, these methods remain rarely used, as evidenced by a recent random sample survey of 57 epidemiological studies that found only one use of quantitative corrections.[5]

All forms of bias may be viewed as arising from missing data.[6] Indeed, treating the bias due to measurement error as a missing-data problem leads to valid and easily employed methods for measurement-error correction [chapter 7 in Ref. (7)]. In the case of exposure measurement error, data are missing on the true exposure for some or all observed participants; only a measurement that is subject to error is observed on everyone in the study. The connection between the mismeasured and true exposure may be constructed either from a subset of data with values for both the mismeasured exposure and a gold standard assumed to equal the true (a validation substudy) or from external information relating the mismeasured exposure to the gold standard. Here, we show how multiple-imputation[8–10] for measurement-error (MIME) correction can be done, based on a validation substudy that is randomly sampled from the total cohort, and we evaluate this approach with a small simulation experiment. This approach (like other validation-based methods) assumes that validation data on subjects are 'missing at random',[8,10] a condition that is satisfied if the validated subjects are a random sample of the

* Corresponding author. Department of Epidemiology, 615 North Wolfe Street—E7640, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA. E-mail: scole@jhsph.edu

observed subjects or a stratified random sample with stratification on observed variables such as sex, age, and disease outcome.

Methods

Hypothetical study population

We base our example and simulations on the expectations of a prospective cohort study of chronic kidney disease currently enrolling 600 children aged 1–16 years with chronic renal insufficiency. A goal of the study is to estimate the effect of low glomerular filtration rate (GFR $<$ 50 ml/min/1.73m^2) at study entry relative to moderate GFR (50–75 ml/min/1.73 m^2) on the 4 year incidence of end-stage renal disease (ESRD). Based on registry data from the North American Pediatric Renal Transplant Cooperative Study,[11] we expect 120 ESRD endpoints over 4 years of follow-up with about a 2-fold hazard comparing the 40% with a low baseline GFR to the remaining 60% with a moderate baseline GFR.

Let T be the minimum of the time from study entry to ESRD or censoring and $D = 1$ if the child gets ESRD during study follow-up, $D = 0$ if the child is censored. Let X be the true GFR status indicator: $X = 1$ if GFR $<$ 50 at study entry, $X = 0$ if GFR $\geqslant$ 50. R indicates inclusion in a random validation substudy: $R = 1$ if the gold-standard GFR status X is observed, $R = 0$ if X is missing. W is the omnipresent but mismeasured version of GFR, with sensitivity Pr ($W = 1 \mid X = 1$) and false-positive rate Pr ($W = 1 \mid X = 0$) = 1 − specificity.

For this hypothetical cohort study the true hazard ratio is 2.25, sensitivity is 0.9, specificity is 0.7, and the size of the validation substudy is 25% of the 600 children. The hypothetical cohort study data comprise values for T, D, W, X, and R (with X missing whenever $R = 0$) for each of 600 children. Appendix 1 gives details on the generation of simulated data.

Analyses of the hypothetical study data

We compared five hazard-ratio estimates for ESRD: (a) comparing those with $W = 1$ with those with $W = 0$, which is termed the naïve analysis; (b) comparing those with $X = 1$ with those with $X = 0$ in the subset with $R = 1$, which is termed the validation substudy (complete-case) analysis; (c) comparing imputed $X = 1$ with imputed $X = 0$ in the whole sample based on MIME; (d) comparing calibrated $X = 1$ with calibrated $X = 0$ in the whole sample based on RC[12]; and (e) an ideal analysis, comparing those with $X = 1$ with those with $X = 0$ in the entire sample, which is termed the complete-data analysis. The last analysis cannot be computed in practice, but serves as a gold standard to evaluate the other analyses, as it is free of the effects of measurement error.

The estimate from MIME correction was obtained by creating 40 completed versions of the observed dataset, where, for those children not in the validation substudy the missing GFR exposure X is replaced with a random draw based on the relation of X and W obtained from those in the validation substudy. The 40 datasets are then combined using standard multiple-imputation techniques. Details are provided in Appendix 2. The key idea to multiple-imputation is to model the distribution of missing data and then create multiple datasets replacing the missing data with random draws from this distribution. Standard complete-data analyses are then conducted on the multiple datasets and combined in a way to reflect the uncertainty regarding the missing data.

The RC estimate was obtained by dividing the estimate from the naïve analysis (a) by the estimated slope from a regression of X on W; we used the variance estimate of Rosner *et al.*[1] SAS version 9 was used for all analyses (SAS Institute Inc, Cary, NC). The SAS code in Appendix 3 carries out the steps needed to implement MIME as described in Appendix 2.

Simulation study

In addition to the example of a single hypothetical cohort study, we present results based on 2000 trials for each of the eight scenarios, one for each combination of the following parameters: percentage in the random validation subset {15%, 25%}, sensitivity {0.7, 0.9}, specificity {0.7, 0.9}. For a given simulation trial, 600 data records were created each composed of a value for T, D, W, X, and R as described in Appendix 1. The eight scenarios were chosen to represent expectations for the prospective cohort study described previously. We also explored relative hazards of 1.5 and 2 under a Weibull distribution with a scale parameter of one (i.e. a constant hazard), sample sizes of 400 and 1000, as well as random validation subsets of 10 and 30%, but present only the eight scenarios described above. The pattern of results observed for the eight scenarios presented was typical of the other scenarios explored. For each scenario, the simulation standard error for the percentage coverage of a valid 95% confidence interval is $100 \times (0.05 \times 0.95/2000)^{1/2} < \frac{1}{2}\%$.

Analyses of simulated data

For each of the 2000 simulation trials across each of the eight scenarios, we estimated the five hazard ratios (a–e) for the risk of ESRD described previously. To compare the estimates, within each scenario we calculated (i) simulated bias, computed as the geometric mean estimate divided by the true hazard ratio minus one then multiplied by 100, (ii) simulated confidence-interval coverage, computed as the proportion of times the confidence interval contains the true hazard ratio, (iii) statistical power, computed as the proportion of simulations in which the 95% confidence interval excludes the null value of one, and (iv) relative square root of the mean squared error or root mean squared error (RMSE), computed as the ratio of the square root of mean squared error for an estimate divided by the square root of mean squared error for the ideal analysis (e). The RMSE was estimated by taking the square root of the sum of the square of the average difference between an estimator and the true log hazard ratio and the simulated variance for the estimator, where the latter is the square of the simulated standard deviation of the estimator. Results are presented for the simulations in which convergence was obtained; 1.3% of the MIME analyses and $<$1% of the validation substudy analyses did not converge. We conducted a separate simulation under the null hypothesis of a hazard ratio of one to ensure that the MIME correction

Table 1 Hypothetical cohort study observed and complete data for 600 children classified by estimated ($W = 1$) and measured ($X = 1$) low GFR at study entry[a]

	Complete data $N = 600$			Validation substudy $N = 150$		
	No. of ESRD events	No. of children	Person-years	No. of ESRD events	No. of children	Person-years
$W = 0$						
$X = 0$	*49*	*252*	*330.0*	13	56	70.6
$X = 1$	*8*	*22*	*24.6*	1	2	2.8
Total	57	274	354.6	14	58	73.4
$W = 1$						
$X = 0$	*23*	*110*	*140.8*	6	35	44.9
$X = 1$	*63*	*216*	*245.2*	20	57	66.7
Total	86	326	386.0	26	92	111.6
Total	143	600	740.6	40	150	185.0

[a] One draw from our simulation in the scenario where sensitivity is 0.9, specificity is 0.7, and the validation substudy is 25%; italicized complete data is typically unobserved.

Table 2 Estimates of the ESRD hazard ratio due to a low GFR at study entry, 95% confidence intervals (95% CLs), and mean squared error based on hypothetical cohort study data for 600 children shown in Table 1[a]

Model	Hazard ratio	95% CL	Mean squared error[b]
(a) Naïve	1.51	1.08, 2.11	0.11
(b) Validation substudy	1.97	1.04, 3.74	0.11
(c) MIME	2.01	1.23, 3.29	0.06
(d) RC	2.01	1.11, 3.65	0.09
(e) Complete data	1.99	1.43, 2.76	0.03

[a] One draw from our simulation in the scenario where sensitivity is 0.9, specificity is 0.7, and the size of the validation substudy is 25%.
[b] Defined as squared bias plus variance, where bias is the difference between the estimated log hazard ratio from each model and that from the complete data.

was valid; with α set at 0.05, 2000 simulation trials yielded a somewhat conservative observed type-I error rate of 0.03 ± 0.005.

Results

Hypothetical cohort study

Table 1 presents the observed and complete data for our hypothetical cohort study under the scenario where the true hazard ratio is 2.25, sensitivity is 0.9, specificity is 0.7, and the size of the validation substudy is 25% of the 600 children. From Table 2, the complete-data hazard ratio for this sample is 1.99 (95% CL 1.43, 2.76) and the naïve (i.e. misclassified) hazard ratio is 1.51 (95% CL 1.08, 2.11).

Of the 600 children, 150 or 25% were included in the validation substudy. A total of 58 of the 274 estimated as unexposed ($W = 0$), and 92 of the 326 estimated as exposed ($W = 1$) were selected at random into the validation substudy. Of the 58 estimated as unexposed 56 were confirmed by measurement as unexposed; while 57 of the 92 estimated as exposed were confirmed by measurement as exposed, yielding observed sensitivity and specificity in the validation substudy of 0.97 ± 0.02 and 0.62 ± 0.05, respectively. Limiting the analysis to the 150 children in the validation substudy yields an estimate of the hazard ratio of 1.97 (95% CL 1.04, 3.74), close to the complete-data value but with a near doubling of the ratio of the upper to lower 95% limits (Table 2).

Applying the MIME correction (Appendix 2) yielded an estimated hazard ratio of 2.01 (95% CL 1.23, 3.29), nearly the same as seen in the complete data. As expected, the interval estimate is narrower (as measured by the ratio of upper to lower 95% limits) than that from the validation substudy alone. The calibration model relating the measured GFR (i.e. X) to the estimated GFR (i.e. W) among the 150 participants in the validation substudy provided an estimate of 0.59 ± 0.07. Therefore, the regression-calibrated hazard ratio is 2.01 = exp [log (1.51)/0.59] with a 95% CL of 1.11, 3.65. In this scenario, MIME provided a narrower interval for the hazard ratio than either restriction to the validation substudy or RC, while all three analyses provided a point estimate close to the complete-data hazard ratio.

Simulations

The MIME correction appeared practically unbiased for every scenario examined, as did the analysis restricted to the random validation substudy and the RC estimator (Table 3). In contrast, the naïve approach (i.e. using the mismeasured GFR) provided an average hazard ratio attenuated towards the null, with the amount of attenuation a function of sensitivity and specificity, as expected.

The confidence limits from MIME correction showed adequate coverage for all scenarios, as did the limits from analyses of the validation substudy alone and from RC (Table 3). Indeed, the slight conservative coverage of the MIME limits coheres with the somewhat conservative type one error rate reported above. In contrast, the limits from the naïve approach exhibited poor coverage, which varied with sensitivity and specificity.

The MIME correction provided uniformly higher statistical power compared with the validation substudy analysis for all scenarios (Table 3). The MIME correction had nearly equal or greater statistical power compared with the naïve analysis and RC when the measurement properties were poor or when the validation substudy was large (i.e. 25% of 600 children).

Combining bias and variance using the RMSE and comparing with the ideal analysis (e), the MIME correction again

Table 3 Bias, 95% CL coverage, and statistical power for a true ESRD hazard ratio of 2.25, 2000 samples of 600 children for each of eight scenarios

		Naïve (a)			Validation substudy (b)			MIME (c)			Regression calibration (d)			Complete data (e)		
Sensitivity, specificity	% In validation substudy	Bias (0)	Cover (95)	Power (99)	Bias (0)	Cover (95)	Power (99)	Bias (0)	Cover (95)	Power (99)	Bias (0)	Cover (95)	Power (99)	Bias (0)	Cover (95)	Power (99)
0.7, 0.7	15	−62	21	41	5	96	40	3	97	46	5	96	27	1	95	99
	25	−61	22	41	2	95	59	1	96	66	5	97	34	0	95	99
0.7, 0.9	15	−38	61	79	2	97	40	3	98	50	−2	96	75	1	95	99
	25	−38	61	79	1	96	59	1	96	73	−2	96	78	0	95	99
0.9, 0.7	15	−40	58	75	3	96	40	1	98	50	6	97	72	0	96	99
	25	−39	60	77	2	96	61	1	97	73	6	97	75	1	95	99
0.9, 0.9	15	−21	84	94	4	95	40	4	99	59	1	97	94	0	95	99
	25	−21	86	95	3	95	62	3	97	85	2	94	95	1	95	99

[a] Bias is defined as 100 × the ratio of the geometric mean of the 2000 estimated hazard ratios divided by the true hazard ratio minus one, percentage coverage is defined as 100 × the proportion of 2000 samples that include the true hazard ratio, and statistical power is defined as 100 × the proportion of 2000 samples that exclude the null hazard ratio of unity; ideal result in parentheses at the top of each column.

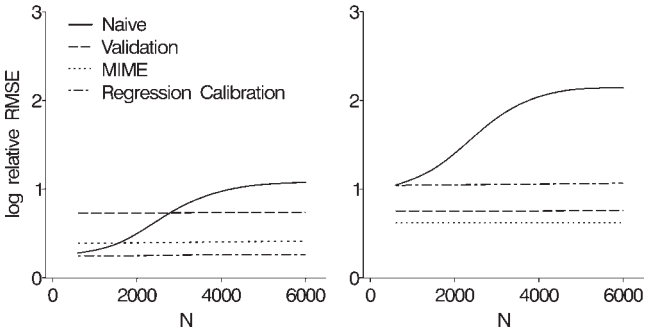

Figure 1 Log of the root mean squared error (RMSE) relative to the complete-data analysis, as a function of sample size (*N*) for two scenarios with a validation substudy of 25% of *N*. On the left, sensitivity and specificity are 0.9. On the right, sensitivity and specificity are 0.7

outperformed analysis restricted to the validation substudy, but again only performed equal to or better than the naïve analysis and RC when measurement properties were poor or when measurement properties were good with a large validation substudy (data not shown). In many of the scenarios explored, the correction for bias was outweighed by added imprecision.

Because the RMSE is largely determined by variance, which is inversely proportional to sample size, we explored the relative RMSE as a function of sample size ranging from 600 to 6000. Results are illustrated in the two panels of Figure 1. As expected, the RMSE for the naïve analysis worsens with increasing sample size (as bias becomes more important relative to variance). With poor measurement (i.e. sensitivity = specificity = 0.7; right panel), MIME appeared to have the lowest RMSE followed by the validation substudy, and then the RC, with the naïve analysis clearly a poor choice regardless of sample size. With good measurement (i.e. sensitivity = specificity = 0.9; left panel), RC appeared to have the lowest RMSE. In this case, the next lowest RMSE after RC depended on sample size. For instance, when sample size was less than ~1500 the naive analysis had the second lowest RMSE. But when sample size was greater than ~1500, the MIME had the second lowest RMSE. At a sample size greater than ~3000, the validation substudy had a lower RMSE than the naïve analysis. The usefulness of the MIME correction is dependent on both the sample size and percentage in the validation substudy, as the validation model becomes unstable with small absolute numbers.

Discussion

In the hypothetical example and other scenarios explored here by simulation, the MIME correction worked well at removing the bias due to non-differential exposure misclassification, provided adequate coverage, and was uniformly more powerful than analysis restricted to the validation substudy (complete-case analysis). The MIME correction was sometimes more powerful than the naïve analysis or RC in the scenarios explored. Furthermore, the MIME approach is general, being based on the flexible approach of multiple-imputation for missing data.[10]

Using the latter approach, MIME can be extended to handle either categorical or continuously distributed mismeasured exposures by appropriately altering the model used to relate the gold standard to the mismeasured exposure in step 1 of Appendix 2, e.g. one could use any generalized linear model to impute true exposure, such as polytomous logistic regression for a polytomous exposure or log-linear regression for a log–normal exposure. For instance, when estimating the effect of GFR on incident ESRD one may wish to specify GFR as a continuous variable rather than an indicator of <50 ml/min/1.73 m^2. Here, we did the latter for pedagogical reasons because we wished to demonstrate MIME in the ubiquitous epidemiological context of a binary exposure and time-to-event. The MIME correction can be applied to disease models other than the proportional-hazards model, such as a log-linear or logistic model, by altering step 4 of Appendix 2. Also, the MIME correction can be used for mismeasured outcome or covariate information provided there are validation data for these variables. In all these applications, MIME can allow for confounders or effect-measure modifiers measured with negligible error (e.g. age and sex) by including covariates and interaction terms in the *K* regression models of step 4 in Appendix 2. Finally, MIME can also allow for arbitrary

missing data, as well as incorporate sensitivity analysis to explore the possible effects of non-random missingness.[9,10] All these forms of MIME correction can be implemented using the SAS procedure MI (SAS Institute Inc, Cary, NC) or the software package SOLAS (Statistical Solutions, Saugus, MA).

If the estimates of the relation between the gold standard and mismeasured variable are highly uncertain, then the MIME correction will propagate this uncertainty by enlarging the estimated variance. The confidence limits for MIME will thus incorporate uncertainty due to both random sampling and measurement errors. As noted in the simulations, when the absolute numbers in the validation substudy are small, the added uncertainty in the MIME may overwhelm the bias reduction and produce RMSE greater than that obtained by the biased, but more precise, naïve estimate. Mean squared error places *equal* weight on the squared bias and variance. The MIME correction will increasingly outperform the naïve analysis as more weight is placed on the squared bias than the variance, as may be the perspective for most epidemiologists. Nonetheless, as with all simulation studies, our simulations cover only a small range of possible scenarios, and it is not always possible to tell which scenario is the best guide to one's own study. While the MIME correction appeared to work well in these simulations, we have not proved the asymptotic consistency of the estimator under the survival-time scenario we examined.

Following standard approaches to imputation,[9,10] we modelled the positive and negative predictive values (i.e. a model for the probability of gold-standard exposure given the mismeasured exposure). One could re-parameterize this problem by modelling the sensitivity, specificity, and marginal X prevalence, and then using Bayes' theorem merge these models into a model for predictive values (which would be highly non-linear). Given non-differential measurement error, when modelling the probability of W given X (sensitivity and specificity) one need not include D as a regressor because in that case W and D are independent given X (here we take D to represent both D and T). In contrast, when modelling the probability of X given W, as done here, one must include D, because one is then modelling predictive values of W for X. Without further assumptions, these predictive values depend on the X prevalence, and X depends on disease given W (except under the null); therefore, in this non-linear case, both D and W are needed to impute X without bias.

When allowing for differential measurement error, however, one must always also include D in the model for W given X. Under differential misclassification, the maximum likelihood (ML) estimates of the predictive values of X given W and D equal the observed proportions with $X = 1$ at $W = w$ and $D = d$.[20] The latter proportions equal the predictive values estimated from step 1 in Appendix 2 by fitting a logistic regression model to the validation substudy and including the interaction term WD. Consequently, under differential misclassification, MIME with an interaction term WD in step 1 of Appendix 2 approaches full efficiency under 'missing at random' and the assumed model as the number of imputations increases. Parameterization based on sensitivity and specificity is more directly linked to the concept of non-differential misclassification; such a parameterization is harder to fit but may provide more direct incorporation of external information. In fact, record-level Monte-Carlo sensitivity analysis for measurement error[13] can be viewed as MIME in which the imputation model in step 1 of Appendix 2 is based entirely on external information instead of validation data.

Our imputation model does not make full use of the non-differential misclassification assumption. If the data are truly subject to non-differential misclassification, then making full use of this assumption would improve the efficiency of our method. To do so, in step 1 of Appendix 2 one could use the ML estimates of the predictive values of X given W and D under non-differential misclassification. Specifically, one could fit a more general log-linear model for the counts of the eight cell table defined by X, W, and D, using ML for misclassified data.[14] Omitting the WD and XWD terms from this log-linear model imposes a non-differential misclassification constraint and yields estimates of predictive values that obey the constraint. Having fit this model, however, one can just use the ML estimates of the XD odds ratio obtained directly from the fitted model,[14] eliminating the need for multiple imputation. In fact, ML estimation for missing data[10] has in theory the same flexibility as multiple imputation, but unfortunately is not as generally implemented in software.

The choice between the MIME correction, RC, ML, or other methods will depend largely upon performance, ease of implementation, and the objectives of analysis. In some scenarios explored, the performance of MIME and RC was similar, but one can find scenarios where either one outperforms the other. For those familiar with multiple-imputation, implementing the MIME correction is straightforward. For those unfamiliar with multiple-imputation, when two imperfect measurements are available rather than an imperfect and a gold standard, or for very large datasets, RC may be easier. Finally, in cases where one wishes to extend the correction to handle differential measurement error or in which multiple-imputation will be done anyway to handle explicitly missing values, MIME has the advantage of requiring little additional programming.

The primary limitation of procedures such as MIME, ML, RC, and external adjustment is that all of them depend on having valid estimates of the relation between the gold standard and mismeasured variable, as would be obtained from an unbiased validation study. Unfortunately, when acquisition of validation information is intrusive, subject refusal may impose unknown degrees of selection bias on the validation study. A potentially worse problem is that the gold-standard variable X itself may be subject to severe and unknown errors, including systematic ones, and that these errors may be correlated across subjects and with errors in the secondary measurement W. Such a situation is only to be expected, for example, when both X and W depend on second-stage data (such as a diet–nutrient table) and errors occur in the latter.[15] In such situations, corrections can even worsen bias and should thus not be taken as more than sensitivity analyses under their assumed models.

In conclusion, viewing bias in epidemiological studies as being due to missing information yields insight into the determinants of bias. This perspective also allows the use of well-developed missing-data methods to account for bias in epidemiological analyses. The use of MIME correction as detailed here provides an example of the practical benefits of this perspective.

Acknowledgements

The authors thank the editor and reviewers for suggestions, which improved this manuscript. Drs Cole and Chu were supported in part by the National Institutes of Health through the data coordinating centres for the paediatric Chronic Kidney Disease cohort study (UO1-DK-066116), Multicenter AIDS Cohort Study (UO1-AI-35043), and the Women's Interagency HIV Study (UO1-AI-42590).

References

[1] Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Stat Med* 1989;**8:**1051–69; discussion 1071–73.

[2] Spiegelman D, Carroll RJ, Kipnis V. Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument. *Stat Med* 2001;**20:**139–60.

[3] Freedman LS, Fainberg V, Kipnis V, Midthune D, Carroll RJ. A new method for dealing with measurement error in explanatory variables of regression models. *Biometrics* 2004;**60:**172–81.

[4] Greenland S. Basic methods for sensitivity analysis of biases. *Int J Epidemiol* 1996;**25:**1107–16.

[5] Jurek A, Maldonado G, Church T, Greenland S. Exposure-measurement error is frequently ignored when interpreting epidemiologic study results. *Am J Epidemiol* 2004;**159:**S72.

[6] Rubin DB. Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics* 1991;**47:**1213–34.

[7] Carroll RJ, Ruppert D, Stefanski LA. *Measurement Error in Nonlinear Models*. New York: Chapman & Hall/CRC, 1995.

[8] Rubin DB. Inference and missing data. *Biometrika* 1976;**63:**581–692.

[9] Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley, 1987.

[10] Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. 2nd edn. New York: John Wiley & Sons, 2002.

[11] North American Pediatric Renal Transplant Cooperative Study (NAPRTCS). *2002 Annual Report*. Rockville, MD: The EMMES Corporation, 2002.

[12] Spiegelman D, Rosner B, Logan R. Estimation and inference for logistic regression with covariate misclassification and measurement error, in main study/validation study designs. *JASA* 2000;**95:**51–61.

[13] Lash TL, Fink AK. Semi-automated sensitivity analysis to assess systematic errors in observational data. *Epidemiology* 2003;**14:**451–58.

[14] Espeland MA, Hui SL. A general approach to analyzing epidemiologic data that contain misclassification errors. *Biometrics* 1987;**43:**1001–12.

[15] Greenland S. The problem of excessive certainty, with special reference to the value of validation studies. *Am J Epidemiol* 2006; (to appear).

[16] Steenland K, Deddens JA. A practical guide to dose–response analyses and risk assessment in occupational epidemiology. *Epidemiology* 2004;**15:**63–70.

[17] Robins JM, Wang NS. Inference for imputation estimators. *Biometrika* 2000;**87:**113–24.

[18] Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *Am J Epidemiol* 1995;**142:**1255–64.

[19] Wang CY, Hsu L, Feng ZD, Prentice RL. Regression calibration in failure time regression. *Biometrics* 1997;**53:**131–45.

[20] Lyles RH. A note on estimating crude odds ratios in case-control studies with differentially misclassified exposure. *Biometrics* 2002;**58:**1034–37.

Appendix 1: Generation of simulated data

Capital letters will represent variables, and lower case letters will represent values of the variables. A simulated data record comprises a value for *T, D, W, X, R*; we draw 600 simulated data records for the hypothetical cohort study as well as for each of the 16 000 simulation datasets. Let U be the time from study entry to ESRD, V the time from study entry to the end of follow-up, $T = \min(U,V)$, and $D = 1$ if $T = U$ (child gets ESRD during study), otherwise $D = 0$ (child is censored). We generated times to ESRD U dependent on GFR according to the Weibull distribution $\alpha\lambda t^{\alpha-1}\exp(-\lambda t^{\alpha})$, where $\lambda = \exp[-1 - \log(1.5) \times X]$ and the Weibull shape parameter $\alpha = 2$, yielding a true summary hazard ratio of 2.25 (= exp [ln (1.5) × 2]) and an increasing hazard over time conditional on GFR. We chose V to yield the expected 80% censored endpoints. W was generated as a Bernoulli variable with probability of $W = 1$ equal to the sensitivity when $X = 1$ and equal to the false–positive rate when $X = 0$. X was set at the expected value of 40%. Poisson versions of the complete-data, naïve, and validation substudy analyses can be directly constructed from the data in Table 1, but the resulting rate ratios underestimate the analogous hazard ratios due to the increasing hazard.

Appendix 2: Multiple imputation for measurement-error (MIME) correction

We implemented the MIME correction as follows.

Step 1: Using standard software, we fit a logistic model relating the gold-standard X to the mismeasured exposure among those subjects in the validation substudy ($R = 1$):

$$\text{logit}\,\Pr(X = 1 | W = w, D = d, T = t) = \alpha_0 + \alpha_1 w + \alpha_2 d + \alpha_3 \ln(t),$$

a model for the predictive values of W as a measure of X. Store the resulting parameter estimates and covariance matrix. Although we did not do so, one could include in this model the product WD, or other observed variables; specifically, absence of WD follows from our assumption of non-differential measurement error (the converse is not true, i.e. absence of WD does not imply non-differential measurement error). To allow for arbitrary differential misclassification one would have to add the product term WD to this model. We assume that log T is log-linearly related to X, but in other settings one may wish to relax the functional form for T by use of a spline or other smoother.[16]

Step 2: Using the estimated parameters $\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3$ and covariance matrix $\hat{\Sigma}_{w,d,t}$ from the fit of the logistic model, draw an estimate of the set of four coefficients for each imputation K from a standard multivariate (4 column) Gaussian distribution. This allows uncertainty about the dependence of X on W, D, and T to propagate through $\hat{\Sigma}_{w,d,t}$.

Step 3: For records in the validation substudy, let $Z_k = X$ for all k, where k indexes the $K = 40$ imputations. Traditionally in multiple-imputation, K is set to a number between 3 and 5.

Since the proportion of missing information was as large as 50% in many of our scenarios and computational resources are inexpensive, we explored the use of K between 10 and 99. $K < 30$ appeared to carry more bias than $K > 30$; so we use $K = 40$ here, which is several times the number of imputations commonly used. For records not in the substudy, draw $Z_k \sim$ Bernoulli $(\hat{p}_{k,w,d,t})$ for $k = 1$ to 40, where $\hat{p}_{k,w,d,t} = \frac{1}{1+exp\left[\hat{\alpha}_0^k+\hat{\alpha}_1^k w+\hat{\alpha}_2^k d+\hat{\alpha}_3^k \ln(t)\right]}$. This step incorporates uncertainty about X for records outside the validation substudy. Steps 2 and 3 together produce the imputed datasets.

Step 4: Fit K analysis models and combine the results using standard multiple-imputation techniques.[10] More explicitly, fit a Cox model $\lambda_{U,k}(t) = \lambda_{0,k}(t) \times e^{\beta_k Z_{ik}}$ for $k = 1$ to K. The estimated hazard ratio is,

$$\exp(\bar{\beta}) = \exp\left(K^{-1}\sum_{k=1}^{K}\hat{\beta}_k\right),$$

where $\hat{\beta}_k$ is the log hazard ratio estimated from the kth imputed dataset in Step 3. The Rubin variance estimate for $\bar{\beta}$ is,

$$V(\bar{\beta}) = K^{-1}\sum_{k=1}^{K}\hat{V}(\hat{\beta}_k) + \left(1+K^{-1}\right)(K-1)^{-1}\sum_{k=1}^{K}\left(\hat{\beta}_k - \bar{\beta}\right)^2,$$

where the first term represents the within-imputation variance and the second term represents the between-imputation variance with a correction for the fact that K is finite.

As does the SAS procedure MI or SOLAS software for multiple-imputation, we used Rubin's simple imputation variance estimator. This estimator can be inconsistent in some cases, and (even when it is consistent) a more complex multiple-imputation approach described by Robins and Wang[17] provides greater accuracy under the assumed imputation model. However, the Robins and Wang variance estimator is computationally more difficult and not available in packaged software.

This imputation scheme uses imputation-coefficient resampling in step 2, and is 'proper' in the sense of Little and Rubin [p. 216 in Ref. (10)]. Omitting the imputation-coefficient resampling as described by Greenland and Finkle[18] and Robins and Wang[17] provides an approximately unbiased and more efficient estimate of the hazard ratio. However, the efficiency gains that are obvious when omitting imputation-coefficient resampling with X and W normally distributed were not apparent with X and W binary in our example. With or without imputation-coefficient resampling, a single imputation will provide an approximately unbiased estimate of the hazard ratio, but multiple imputations are needed to correctly reflect the uncertainly due to measurement error.

RC can be derived as an instrumental-variable estimator or as a single-imputation from the model for X on W; assuming non-differential error, the former formulation divides the naïve estimate (which is coefficient for the dependence of D on W) by the estimated dependence of X on W. Further research is needed to elucidate why RC formulated as a single-imputation based on the model for X on W, and those based on a model for X on W and D as is used here for MIME, both appear to yield consistent estimates of the measurement error corrected association between X and D. Finally, in our simulations, we did not notice the bias in RC that has been attributed to the dilution of the risk sets over follow-up time.[19]

Appendix 3: SAS code to carry out a simple MIME correction as described in Appendix 2

```
*step 1 – Fit validation model, save betas and cov matrix;
proc logistic data = a descending covout outest = b(keep = _name_ intercept w d lnt) noprint;
  where r = 1; model x = w d lnt;
data beta(keep = intercept w d lnt); set b; if _name_ = ''x'';
data cov(keep = intercept w d lnt); set b; if _name_ = ''x'';

*step 2 – Imputation-coefficient resampling;
proc iml;
use cov; read all into cov;                  *variance-covariance matrix;
  use beta; read all into mu; mu = mu';      *means;
  v = nrow(cov);                             *number of variables;
  n = 40;                                    *number of imputations;
  seed = 372;
  l = t(root(cov));                          *cholesky root of cov matrix;
  z = normal(j(v,n,seed));                   *generate nvars*samplesize normals;
  x = l*z;                                   *premultiply by cholesky root;
  x = repeat(mu,1,n)+x;                      *add in the means;
  tx = t(x);
  create m from tx; append from tx; *write out sample data to sas dataset;
quit;
data m(keep = _imputation_ b0-b3); set m; retain _imputation_ 0; _imputation_ = _imputation_+1;
  b0 = col1; b1 = col2; b2 = col3; b3 = col4;
data aa; set a; do _imputation_ = 1 to 40; output; end;

*step 3 – Generate imputations;
proc sort data = aa; by _imputation_;
```

```
proc sort data = m; by _imputation_;
data c; merge aa m; by _imputation_; call streaminit(320);
  if r = 1 then v = x; else v = rand(''bernoulli'',1/(1+exp(-(b0+b1*w+b2*d+b3*lnt))));

*step 4 – Fit Cox models and combine information across imputations;
proc phreg data = c noprint outest = e covout; model t*d(0) = v; by _imputation_;
proc mianalyze data = e; modeleffects v; title ''mime'';
```

Published by Oxford University Press on behalf of the International Epidemiological Association *International Journal of Epidemiology* 2006;**35:**1081–1082

doi:10.1093/ije/dyl139

Commentary: Dealing with measurement error: multiple imputation or regression calibration?

Ian R White

Multiple imputation (MI) is a well-established method of handling missing data and is increasingly implemented in statistical software packages. Unlike other imputation methods, MI produces not one but several imputed datasets. This enables it to appropriately reflect the uncertainty due to missing data and, hence, to produce valid statistical inferences.[1]

Cole *et al*. in this issue[2] propose that MI may also be useful in dealing with a second problem rife in epidemiology: exposure measurement error, which typically causes underestimation of exposure–disease associations (regression dilution bias).[3] They coin the acronym MIME (multiple imputation for measurement error) and show that this method can indeed remove regression dilution bias. How widely should MIME be used?

Unfortunately, MIME is only appropriate for measurement error problems in which the true exposure is measured in a sub-sample (a validation study). This is because MIME involves fitting a regression model of true exposures on observed exposures, in order to impute the unobserved true exposures. Often, the degree of measurement error is assessed by taking repeat measurements (a repeatability study).[4] In such cases, the true exposure is never observed, so MIME as described by Cole *et al*. would not be appropriate (and complex modifications would be required to make MIME work).

The main alternative to MIME is regression calibration (RC).[5] RC replaces observed exposures by predicted values of the true exposure; using these predicted values as if they were the true exposure can yield valid estimates of the true exposure–disease relationship.[6] Both MIME and RC are based on regression models for the true exposure (although disease status is included in the model for MIME and not for RC). But whereas MIME must create several possible values of true exposure, RC creates only a single predicted value: this is why RC works with a repeatability study.

With validation studies, is MIME superior to the RC approach? Cole *et al*. report in their abstract that MIME 'was sometimes more powerful than misclassified or RC analyses'. Although this is true, it is fair to point out that MIME was the most powerful of the three analyses in only two of eight scenarios considered (Table 3); in the other six scenarios it was the least powerful of the three. Surprisingly, Cole *et al*. also found that RC was uniformly less powerful than misclassified analyses: however, the difference in power would have disappeared if they had used asymmetric confidence intervals based on Feiller's theorem[7] instead of symmetrical intervals based on Rosner's variance estimate.

There is one good theoretical reason for expecting MIME to perform better than RC. MIME uses the true exposure when it is available, rather than imputing a value, whereas RC always predicts the true exposure from the observed exposure. It is, therefore, surprising that MIME performed so badly in several cases in Cole *et al*.'s simulation studies. Possible explanations include the difficulty in performing multiple imputation with survival outcomes[8] and extreme estimates in a minority of imputed datasets.

Finally, it is important to remember that the main problem in epidemiology is measurement error in confounders, not in exposures.[9] Whereas measurement error in exposures dilutes exposure–disease associations, measurement error in confounders can lead to overestimation of associations. Fortunately, given adequate information about measurement error, both MIME and RC can be directly extended to handle measurement error in confounders.[10]

MRC Biostatistics Unit, Institute of Public Health, Cambridge CB2 2SR, UK. E-mail: ian.white@mrc-bsu.cam.ac.uk

References

[1] Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley and Sons, 1987.