# Evaluating Public Health Interventions: 4. The Nurses' Health Study and Methods for Eliminating Bias Attributable to Measurement Error and Misclassification

The Nurses' Health Study and many other large longitudinal cohorts around the world use the food frequency questionnaire to assess dietary intake over time, and to relate diet to health.

Controversies concerning this questionnaire's ability to adequately measure diet have led to a flurry of methods for evaluating the magnitude of measurement error and misclassification in exposure assessment, and for correcting the point and interval estimates of effect on the basis of these assessment methods for this error. Nurses' Health Study investigators have been in the forefront of these developments and their applications, although hundreds of other investigators have also used them.

This commentary provides an overview of the methods and their uses, and concludes with remarks on their potential applications in the evaluation of public health interventions. (*Am J Public Health.* 2016;106: 1563–1566. doi:10.2105/AJPH.2016.303377)

*Donna Spiegelman, ScD*

In honor of this issue of *AJPH* commemorating the 40th anniversary of the launch of the Nurses' Health Study (NHS), I highlight the relevance for public health interventions of methods for eliminating bias attributable to measurement error and misclassification, emphasizing those that have been developed and disseminated by NHS investigators.

I have been deeply involved in these activities as the statistician for the Health Professionals Follow-Up Study since 1992 and for the NHS II since 1994, while playing a statistical supporting role in the NHS since I began my thesis research at the Harvard School of Public Health in 1987. In 1986, after passing my qualifying examinations in the Biostatistics Department and then the Epidemiology Department, I was exploring ideas for a dissertation topic that satisfied the requirements of both departments. I met with Walter Willett, MD, DrPH, from whom I had taken a course on nutritional epidemiology that I found very interesting. He suggested measurement error. The suggestion instantly resonated—I never considered anything else. This led to a more than 25-year career of independent statistical research on the development of methods for the design and analysis of epidemiological studies with correction for bias attributable to exposure measurement error and misclassification.

As taught in modern epidemiology,[1] the three major sources of bias in observational research are confounding, information bias, and selection bias. Epidemiological methods focus on approaches for reducing, if not eliminating, them. Blair et al. wrote in 2007,

> We believe of the two of the major methodological issues raised in epidemiologic studies of occupational exposures, that is, confounding and exposure misclassification, the latter is of far greater concern.[2(p205)]

They continued,

> It is rare to find substantial confounding in occupational studies (or in other epidemiologic studies for that matter), even by risk factors that are strongly related to the outcome of interest. On the other hand, exposure misclassification probably occurs in nearly every epidemiologic study.[2(p205)]

Reviews of the effects of exposure measurement error and the numerous methods that have been proposed to correct for biases that result when exposure measurement error is present have been published.[3]

There have been many theoretical investigations of the effects of measurement error on point and interval estimates of exposure–disease associations, especially focused upon the widespread and profound presence of exposure measurement error and misclassification in environmental and nutritional epidemiology.[4] Nevertheless, until the early 1990s, few original scientific publications made use of methods for empirical correction of bias attributable to exposure measurement error in epidemiology. In fact, the NHS and some of its principal scientific investigators, including Walter Willett, Bernard Rosner, and myself, played a pivotal role in the development and widespread adoption of these methods and the further growth of this statistical field.

In Appendix A (available as a supplement to the online version of this article at http://www.ajph.org), I list additional references for each of the topics

**ABOUT THE AUTHOR**

*Donna Spiegelman is with departments of Epidemiology, Biostatistics, Nutrition, and Global Health, Harvard T. H. Chan School of Public Health, Boston, MA.*

*Correspondence can be sent to Donna Spiegelman, Harvard T. H. Chan School of Public Health, 677 Huntington Ave, Kresge Building Room 802, Boston, MA 02115 (e-mail: stdls@hsph.harvard.edu). Reprints can be ordered at http://www.ajph.org by clicking the "Reprints" link.*

*This article was accepted July 4, 2016.*

*doi: 10.2105/AJPH.2016.303377*

discussed in this article for readers interested in exploring them in more depth.

## EXPOSURE MEASUREMENT ERROR CORRECTION

After I chose measurement error as the focus of my thesis research, Willett asked me to take a look at a paper he and Rosner had submitted to, and that had been rejected by a leading epidemiology journal that proposed a simple, intuitive method for correcting for bias attributable to measurement error in logistic regression. Willett had become intensely interested in the topic of measurement error in nutritional epidemiology in response, in part, to the claim that the reason the NHS failed to find an association between dietary fat intake and breast cancer, as was suggested in 1975 by international ecologic correlation data,[5] was because of the measurement error in the ascertainment of dietary fat intake.[6] The NHS used, and still uses, food frequency questionnaires (FFQs) to track dietary intake over time among the more than 100 000 participants every four years. With my dual training in epidemiology and biostatistics, I was able to provide theoretical justification for the method and conducted an extensive simulation study that demonstrated its excellent finite sample properties, and the article was published.[7]

The FFQ has been validated on the basis of several weighed dietary records. $Z$ stands for the nutrient (e.g., fat) or the food item (e.g., meat) measured with the FFQ and $X$ for the same nutrient or food item measured with the dietary records. When it is reasonable to model the conditional mean for the true exposure, denoted by $X$, given the surrogate, denoted by $Z$, by a linear model, that is

$$(1)\quad E(X|Z) = \gamma_0 + \gamma_1 Z,$$

then the estimated coefficient for the effect of the exposure on the outcome from the logistic model $\beta_1$ (i.e., the log relative risk of disease in relation to the nutrient or food intake),

$$(2)\quad \mathit{logit}[\Pr(D=1)] = \beta_0 + \beta_1 Z,$$

can be corrected for bias attributable to exposure measurement error by

$$(3)\quad \hat{\beta}_1^{\star} = \frac{\hat{\beta}_1}{\hat{\gamma}_1}.$$

Because $\gamma_1$ is usually less than 1, measurement error correction usually de-attenuates the log relative risk, $\hat{\beta}_1$.

Subsequently, we and others[8,9] presented fully multivariate versions of this method, which has become known as the *regression calibration* method. The method permits approximately unbiased point and interval estimates of effect from linear, Cox, and logistic regression models when the usual methods of exposure assessment can be validated against an unbiased but possibly imperfect gold standard[10] in a subsample, which can be within the original main study or outside it. This method has been used in hundreds of original scientific publications (see Appendix B, available as a supplement to the online version of this article at http://www.ajph.org). One reason these methods have been widely used is because of the availability of publicly available, user-friendly software (see http://www.hsph.harvard.edu/donna-spiegelman/software).

## EXTENSIONS MOTIVATED BY THE NHS

Concerns were raised about the validity of the assumptions made by the original regression calibration method. Of particular concern was the "correlated errors" problem, which may result when one self-reported measure is used to validate another.[11] For example, study participants who believe that higher dietary fat intake is unhealthy may underreport higher-fat foods on both the FFQ and their weighed diet records. This led to a series of articles by myself, Rosner and Willett, and other colleagues proposing augmented study designs and extensions to the original method to address these concerns.[12] Even Rosner's daughter, Sarah Rosner Preis, ScD, MPH, now an assistant professor of biostatistics at the Boston University School of Public Health, got involved.[13]

For example, polyunsaturated fat intake (% of total energy intake) has been found in NHS to be protective against diabetes risk.[14] In analysis uncorrected for measurement error in polyunsaturated fat intake, the relative risk (RR) was 0.74 (95% confidence interval [CI] = 0.66, 0.84) for a 12% increase in percentage of calories from polyunsaturated fat, corresponding to an increase from the 10th to 90th percentile of the distribution in the cohort. When possible correlated errors between the FFQ and the four one-week weighed food records used to validate it were ignored, the standard measurement error approach gave an RR of 0.45 (95% CI = 0.28, 0.72) for the same change in intake, and the extended method taking correlated errors into account gave a very similar result, RR = 0.42 (95% CI = 0.27, 0.64), suggesting that the correlated errors problem may be less of a concern than had been suggested.

Further extensions to regression calibration have been developed to take advantage of the unbiased estimates of the exposure effect available in validation studies with sufficient outcome data; situations in which the exposure distribution is highly skewed, as often occurs with micronutrient data and many environmental exposures, leading to heteroscedasticity in the measurement error model variance; cases in which there are multiple surrogates for the same underlying exposure of interest, as commonly arises in environmental health studies; and for studies in which continuous exposures and mismeasured and categorical exposures are misclassified (Appendix A).

## FOOD FREQUENCY QUESTIONNAIRE VALIDATION STUDIES

To empirically adjust point and interval estimates for measurement error, validation and reliability studies are needed; otherwise, it is impossible to fit Equation 1 to estimate $\gamma_1$ and perform the correction. Chapter 6 of Willett's textbook, *Nutritional Epidemiology*[15] cites approximately 200 or more validation and reliability studies from around the world that quantitatively assess the extent of error in a wide range of self-reported and measured variables relevant to epidemiological research, including foods, nutrients, physical activity, serum hormones, and many others. Statisticians, myself included, have developed methods to guide the design of these studies;

in fact, one of my four thesis articles addressed this topic.[16]

Most recently, NHS investigators have completed the Women's Lifestyle Validation Study, which includes measurements of multiple types of repeated objective and self-reported dietary and physical activity assessments, from nearly 800 women, all of whom are members of NHS I or NHS II. This validation study will clarify the validity of total energy intake, protein intake, protein density, sodium intake, potassium intake, and physical activity as measured by FFQs, diet records, and 24-hour online recalls, by comparison with unbiased biomarkers. By using new methods developed by our group,[17] it will soon be possible to estimate the correlations between the errors in sets of self-reported measures and to assess the validity of new technology, such as accelerometry, for assessing physical activity.

## MISCLASSIFICATION CORRECTIONS FOR SURVIVAL DATA

Although much of the initial methodological work in the epidemiology and biostatistics literature has focused on correction for bias attributable to measurement error and misclassification in two-by-two tables, logistic regression, and generalized linear models, in fact, for the past 20 years or more, the primary analytic model used in NHS is the Cox regression model for survival analysis. This model estimates incidence rate ratios (often termed hazard ratios in the statistical literature) and has the attractive feature that it makes no assumptions about the shape of the incidence rate curve as it changes over "time," typically age in the chronic disease epidemiology focus of NHS. In addition, the Cox model is well suited for several additional salient features of NHS: it readily allows for staggered entry of participants into the cohort according to their age at enrollment in 1976, ranging from 29 to 55 years, and it naturally allows for updating of diet every four years and two years for nearly everything else. Although loss to follow-up is rare given the high follow-up rates of the participants even 40 years past enrollment, the Cox model also seamlessly incorporates this feature.

With recent interest in factors promoting healthy aging, especially following cancer diagnoses, the ability of the Cox model to validly estimate rate ratios in the presence of high mortality from the cause of interest as well as other causes has become a more important advantage. More work was thus needed to develop methods for bias correction attributable to measurement error and misclassification appropriate for such censored survival settings with time-varying covariates. After laying down some foundations in early work by myself et al. at Harvard[9] and, separately, at the Fred Hutchinson Cancer Research Center for survival data analysis with baseline covariates only,[18] we drilled down to the problems most relevant to NHS, first addressing bias correction methods for simple time-varying covariates.[19] Now, in our most recent work, we have addressed methods for mismeasured time-varying exposure metrics that are functions of the entire mismeasured exposure history,[20] such as the cumulatively updated average, the primary exposure variable in most NHS dietary analyses. By using these methods, a recent article revised estimates of the impact of the air pollutant, particulate matter with a diameter smaller than 2.5 microns (PM2.5), on all-cause mortality in NHS, finding a nearly double relative risk estimate after measurement error correction.[21] As always, publicly available, user-friendly software is available that implements this method for cumulatively updated averages, cumulative totals, simple updates, and baseline covariates (https://www.hsph.harvard.edu/donna-spiegelman/software/rrc-macro).

What does the future hold for further advances in the development of methods to address exposure measurement error in the NHS? We are looking at issues arising in life-course epidemiology for the estimation of the start and end of time windows of susceptibility in the presence of exposure measurement error, for adjusting the population attributable risk for exposure misclassification, and there will be more to come.

## RELEVANCE TO PUBLIC HEALTH INTERVENTIONS

Bringing it back home, so to speak, I conclude with some remarks about the relevance of methods for the correction of bias attributable to measurement error and misclassification that have arisen out of the NHS to implementation science and the evaluation of large-scale public health interventions. For example, in evaluations of randomized interventions under the intent-to-treat principle, exposure misclassification is eliminated by design. However, even in randomized evaluations, when adherence, fidelity, and compliance are to be accounted for or are of interest, as they typically are, much of the methodology described previously is useful. As always, empirical validation of the measures of adherence, fidelity, and compliance in a subsample is required. Examples of such applications need to be developed and published.

Next, in cluster-randomized and observational interventions, as discussed in previous columns in this series, confounding is always a concern. In interventions at scale, typically making use of routine administrative records in place of often prohibitively costly research-quality data collection procedures, mismeasurement of confounders will lead to residual confounding of the intervention effect if bias correction methods as described previously are not applied. Again, empirical validation of the extent of error in the key confounders is required in a subsample, and in many situations, external validation data from elsewhere may be reasonably be considered transportable, making it possible to produce less-biased results than if confounder measurement error were ignored.

At least one other potential application of methods for bias correction attributable to misclassification and measurement error in implementation science and related disciplines concerns outcome misclassification in the "big data" setting of investigations nested within large administrative public health databases—for example, in comparative effectiveness research. Here, the "phenotypes" are created through a complex process making use of *International Classification of Diseases* codes, natural language processing[22] of textual information notated by the providers, pharmacy records, and other sources. These phenotypic algorithms may be

subject to substantial misclassification. For example, algorithms for identifying the diabetes "phenotype" within electronic medical records have been reported to have quite a large amount of uncertainty.[23] The resulting misclassification, usually of the outcomes in subsequent electronic medical records–based analysis, will lead to substantial bias in effect estimates,[24] unless the definitions are validated and the validation data are used for bias correction.[25]

The past 40 years of the NHS have stimulated an abundance of methodological research, mostly, but not exclusively, in the area of bias correction for exposure measurement error and misclassification. Attributable in large part to the public availability of user-friendly software, the methodology has been used by epidemiological investigators around the world. The methods have a number of potentially important applications in implementation science, comparative effectiveness research, and impact and program evaluations, and it is my hope that they will increasingly be applied in these areas. As always, to use these methods, key mismeasured variables must be empirically validated in relatively small subsamples internal or external to the primary data source. AJPH

## ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health (grant DP1ES025459), the Nurses' Health Study, and the Health Professionals Follow-Up Study.

## REFERENCES

1. Rothman KJ. *Modern Epidemiology*. 1st ed. Boston, MA: Little, Brown; 1986.

2. Blair A, Stewart P, Lubin JH, Forastiere F. Methodological issues regarding confounding and exposure misclassification in epidemiological studies of occupational exposures. *Am J Ind Med*. 2007;50(3):199–207.

3. Armstrong BG. The effects of measurement errors on relative risk regressions. *Am J Epidemiol*. 1990;132(6):1176–1184.

4. Willett WC, Sampson L, Stampfer MJ, et al. Reproducibility and validity of a semiquantitative food frequency questionnaire. *Am J Epidemiol*. 1985;122(1):51–65.

5. Armstrong B, Doll R. Environmental factors and cancer incidence and mortality in different countries, with special reference to dietary practices. *Int J Cancer*. 1975;15(4):617–631.

6. Prentice RL, Kakar F, Hursting S, Sheppard L, Klein R, Kushi LH. Aspects of the rationale for the Womens Health Trial. *J Natl Cancer Inst*. 1988;80(11):802–814.

7. 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(9):1051–1069, discussion 71–73.

8. Carroll RJ, Stefanski LA. Approximate quasi-liklihood estimation in models with surrogate predictors. *J Am Stat Assoc*. 1990;85:652–663.

9. Spiegelman D, McDermott A, Rosner B. Regression calibration method for correcting measurement-error bias in nutritional epidemiology. *Am J Clin Nutr*. 1997;65(suppl 4):1179S–1186S.

10. Spiegelman D, Schneeweiss S, McDermott A. Measurement error correction for logistic regression models with an "alloyed gold standard." *Am J Epidemiol*. 1997;145(2):184–196.

11. Kipnis V, Subar AF, Midthune D, et al. Structure of dietary measurement error: results of the OPEN biomarker study. *Am J Epidemiol*. 2003;158(1):14–21, discussion 2–6.

12. Spiegelman D, Zhao B, Kim J. Correlated errors in biased surrogates: study designs and methods for measurement error correction. *Stat Med*. 2005;24(11):1657–1682.

13. Preis SR, Spiegelman D, Zhao B, Moshfegh A, Baer DJ, Willett WC. Random and correlated errors in gold standards used in nutritional epidemiology: implications for validation studies. *Am J Epidemiol*. 2011;173(6):683–694.

14. Salmerón J, Hu FB, Manson JE, et al. Dietary fat intake and risk of type 2 diabetes in women. *Am J Clin Nutr*. 2001;73(6):1019–1026.

15. Willett W, Lenart E. Reproducibility and validity of food-frequency questionnaires. In: Willett W. *Nutritional Epidemiology*. 3rd ed. Oxford, England: Oxford University Press; 2013.

16. Spiegelman D, Gray R. Cost-efficient study designs for binary response data with Gaussian covariate measurement error. *Biometrics*. 1991;47(3):851–869.

17. Spiegelman D, Pollack S, Chomistek A, Yuan C, Rimm E, Willett W. Generalized methods-of-moments estimation and inference for the assessment of multiple imperfect measures of diet and physical activity in validation studies. Poster presented at: World Epidemiology Congress; June 21–24, 2016; Miami, FL.

18. Xie SX, Wang CY, Prentice RL. A risk set calibration method for failure time regression by using a covariate reliability sample. *J R Stat Soc B*. 2001;63:855–870.

19. Liao X, Zucker DM, Li Y, Spiegelman D. Survival analysis with error-prone time-varying covariates: a risk set calibration approach. *Biometrics*. 2011;67(1):50–58.

20. Liao X, Wang M, Hart J, Laden F, Spiegelman D. Survival analysis with functions of mis-measured covariate histories: the case of chronic air pollution exposure in relation to mortality in the Nurses' Health Study. 2015. Harvard University Biostatistics Working Paper 198.

21. Hart JE, Liao X, Hong B, et al. The association of long-term exposure to PM2.5 on all-cause mortality in the Nurses' Health Study and the impact of measurement-error correction. *Environ Health*. 2015;14(1):38.

22. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *J Am Med Inform Assoc*. 2011;18(5):540–543.

23. Richesson RL, Rusincovitch SA, Wixted D, et al. A comparison of phenotype definitions for diabetes mellitus. *J Am Med Inform Assoc*. 2013;20(e2):e319–e326.

24. Copeland KT, Checkoway H, McMichael AJ, Holbrook RH. Bias due to misclassification in the estimation of relative risk. *Am J Epidemiol*. 1977;105(5):488–495.

25. Lyles RH, Tang L, Superak HM, et al. Validation data-based adjustments for outcome misclassification in logistic regression: an illustration. *Epidemiology*. 2011;22(4):589–597.

## EDITOR'S NOTE

Because of space restrictions and the large volume of references relevant to the Nurses' Health Study, additional references are provided in a supplement to the online version of this article at http://www.ajph.org.