Journal of Exposure Science and Environmental Epidemiology (2012) 22, 409–416
© 2012 Nature America, Inc. All rights reserved 1559-0631/12
www.nature.com/jes

ORIGINAL ARTICLE

Using multiple imputation to assign pesticide use for non-responders in the follow-up questionnaire in the Agricultural Health Study

Sonya L. Heltshe[1,2], Jay H. Lubin[1], Stella Koutros[1], Joseph B. Coble[1], Bu-Tian Ji[1], Michael C.R. Alavanja[1], Aaron Blair[1], Dale P. Sandler[3], Cynthia J. Hines[4], Kent W. Thomas[5], Joseph Barker[6], Gabriella Andreotti[1], Jane A. Hoppin[3] and Laura E. Beane Freeman[1]

The Agricultural Health Study (AHS), a large prospective cohort, was designed to elucidate associations between pesticide use and other agricultural exposures and health outcomes. The cohort includes 57,310 pesticide applicators who were enrolled between 1993 and 1997 in Iowa and North Carolina. A follow-up questionnaire administered 5 years later was completed by 36,342 (63%) of the original participants. Missing pesticide use information from participants who did not complete the second questionnaire impedes both long-term pesticide exposure estimation and statistical inference of risk for health outcomes. Logistic regression and stratified sampling were used to impute key variables related to the use of specific pesticides for 20,968 applicators who did not complete the second questionnaire. To assess the imputation procedure, a 20% random sample of participants was withheld for comparison. The observed and imputed prevalence of any pesticide use in the holdout dataset were 85.7% and 85.3%, respectively. The distribution of prevalence and days/year of use for specific pesticides were similar across observed and imputed in the holdout sample. When appropriately implemented, multiple imputation can reduce bias and increase precision and can be more valid than other missing data approaches.

Journal of Exposure Science and Environmental Epidemiology (2012) **22,** 409–416; doi:10.1038/jes.2012.31; published online 9 May 2012

Keywords: agriculture; cohort studies; missing data; pesticides; precision

INTRODUCTION

Missing data is a common problem in epidemiological studies and the statistical implications of ignoring missing data are well known, including loss of statistical power and potentially biased estimates of association. The multiple imputation technique[1] is an approach whereby the investigator replaces each missing value with several plausible values sampled from a probability distribution, conducts multiple analyses for replicate datasets built from each plausible value, then combines the multiple results to account for the fact that the replacement data were imputed. Multiple imputation has been widely accepted and has been used to account for missing data in large national surveys and studies, including NHANES III,[2] National Assessment of Educational Progress,[3] Children's Mental Health Initiative,[4] and the Framingham Heart Study;[5] however, detailed accounts of the application of multiple imputation and particularly the evaluation and validation of the methods are not often published. This paper demonstrates a practical implementation of multiple imputation and is vital for investigators of the Agricultural Health Study (AHS).

The AHS is a prospective cohort study designed to evaluate the effect of agriculturally related exposures on health outcomes. The study includes 57,310 licensed pesticide applicators from Iowa and North Carolina, as well as 32,345 spouses of licensed applicators, who are not included in this imputation. In Iowa, both private applicators, who are primarily farmers, and commercial applicators were included. In North Carolina, only private applicators were enrolled. Cancer incidence and mortality are obtained by annual linkage to state cancer and mortality registries and to the National Death Index. Exposure information is collected by questionnaire. In the Phase 1 enrollment period (1993–97), applicators provided information on the use of 50 specific pesticides through completion of two self-administered questionnaires that included information on demographics, health history, and lifetime farming and pesticide use practices.[6–8] The study was approved by the Institutional Review Boards of the National Institutes of Health (Bethesda, Maryland) and its contractors. From the enrollment data, two exposure metrics were developed; the first was lifetime days of pesticide use, calculated as the product of years of use of each specific pesticide and average number of days used per year. The second metric, intensity-weighted lifetime days of use, incorporated information about factors that might impact exposure, such as the use of personal protective equipment, whether the applicator mixed pesticides, performed equipment repair, and methods of application.[9] Five years later in Phase 2 (1999–2005), we administered a computer-assisted telephone interview questionnaire that described pesticide use since enrollment. Specifically,

[1]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Rockville, Maryland, USA; [2]Department of Pediatrics, School of Medicine, University of Washington, Seattle, Washington, USA; [3]Epidemiology Branch, National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services, Research Triangle Park, North Carolina, USA; [4]National Institute for Occupational Safety and Health, Cincinnati, Ohio, USA; [5]National Exposure Research Laboratory, United States Environmental Protection Agency, Research Triangle Park, North Carolina, USA; [6]Information Management Systems, Rockville, Maryland USA. Correspondence to: Dr. Laura E. Beane Freeman, Occupational and Environmental Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, 6120 Executive Boulevard, RM 8112, MSC 7240, Rockville, Maryland, 20892, USA.
Tel.: +301 451 8793. Fax: +301 402 1819.
E-mail: freemala@mail.nih.gov
Received 22 September 2011; accepted 27 January 2012; published online 9 May 2012

participants were asked about the last year that they applied pesticides, which was denoted as the Phase 2 reference year, and the type and frequency of use of specific pesticides. A total of 36,342 (63%) of the original participants completed the questionnaire; 8% had died between enrollment and the administration of Phase 2, 15% refused, and 14% could not be reached.[10] For epidemiological analyses, pesticide use information collected in Phase 2 was cumulatively added to information collected in Phase 1 for both aforementioned exposure metrics, using details of specific pesticide use.

When using pesticide exposure in an analysis, there are several ways to handle missing Phase 2 information, including omission of those subjects, simple imputation (e.g., mean value substitution), or ignoring non-response in Phase 2 and implicitly assume zero pesticide exposure after Phase 1, which would be erroneous for most participants who did not complete the Phase 2 questionnaire. To correct for this potential bias, a data-driven multiple imputation for the 20,968 applicators (37%) who did not complete the Phase 2 questionnaire was employed. This paper describes the complex, multi-step process used to impute missing information on pesticide use from Phase 2 and an evaluation of the imputation procedure based on a holdout subset of participants with complete data (i.e., individuals who completed both Phase 1 and Phase 2). We also discuss the assumptions and advantages of multiple imputations.

MATERIALS AND METHODS

Imputation Strategy

An overarching principal of multiple imputation is to model the response of interest, in this case the use of pesticides in the interim period between the administration of the Phases 1 and 2 questionnaires. We used covariates from participants with complete data from both phases, and then applied that model to participants missing Phase 2 to obtain estimates of the missing data. Our specific multiple imputation procedure imputes four primary AHS exposure metric variables of interest: (1) use (yes/no) of any pesticide in the interim period between Phases 1 and 2; (2) use (yes/no) of 50 specific pesticides in the interim period (see Table 1); (3) number of days of use for a specific pesticide during Phase 2; and (4) last year of application of any pesticides within the 5-year period between Phases 1 and 2 (Phase 2 reference year). Phase 2 respondents report use of many pesticides that were not specifically on the Phase 1 questionnaires; however, we limit this imputation to the subset of 50 pesticides that were chosen as the focus in Phase 1. The value of days of use per year on the Phase 2 questionnaire is a discrete count variable that was collapsed into categories and therefore skewed, and reference year is an ordinal variable. We use logistic regression and stratified sampling to impute the 102 variables (any use of pesticides: reference year of use, and for 50 specific pesticides: any use, and days per year) from Phase 2 that are needed to construct the pesticide-exposure metrics in the AHS.

We withheld a randomly selected subset (20%, $n = 7269$) of participants from both Phase 1 and Phase 2 data to assess the proposed imputation method. We compared true and imputed percent usage and days/year of pesticide use within this subset using graphical displays and calculated the Brier score and Brier skill score[11–13] — measures of prediction accuracy. After assessment, the complete data were used to generate the final imputed datasets; nothing was withheld. All analyses were based on AHS data releases P1REL201005.00 and P2REL201007.00 and performed using SAS Version 9.1.

Table 1. Phase 2 (1999–2005) pesticide usage in the AHS: observed and imputed.

	Prevalence estimates (%)		
	Observed (N = 36,342)	*Imputed[a] (N = 20,968)*	*Observed and imputed[a] (N = 57,310)*
Personally mix/load/apply any pesticides	85.21	82.82	84.33
METHYL BROMIDE	0.51	0.49	0.51
ALUMINUM PHOSPHIDE	0.79	0.84	0.81
CARBON TETRACHLORIDE/ DISULFIDE	0.00	0.00	0.00
ETHYLENE-DIBROMIDE	0.03	0.02	0.03
BENOMYL	0.40	0.30	0.36
CHLOROTHALONIL	2.53	2.75	2.61
CAPTAN	2.37	1.65	2.11
MANEB/MANCOZEB	0.18	0.14	0.16
METALAXYL	2.52	2.60	2.55
ZIRAM	0.10	0.08	0.10
ATRAZINE	31.16	25.86	29.22
DICAMBA	19.35	15.31	17.87
CYANAZINE	1.64	1.44	1.57
CHLORIMURON-ETHYL	3.24	3.19	3.22
METOLACHLOR	14.74	13.03	14.11
EPTC	0.35	0.30	0.33
ALACHLOR	2.81	2.49	2.69
METRIBUZIN	1.96	1.62	1.84
PARAQUAT	2.08	2.19	2.12
PETROLEUM OIL/PETROL. DISTILLATES	0.58	0.41	0.52
PENDIMETHALIN	11.71	10.77	11.37
IMAZETHAPYR	8.16	6.68	7.62
GLYPHOSATE	51.82	43.98	48.95
SILVEX	0.00	0.00	0.00
BUTYLATE	0.09	0.08	0.09
TRIFLURALIN	11.10	9.13	10.38
2,4-D	37.32	29.54	34.47
2,4,5-T	0.14	0.11	0.13
PERMETHRIN (for crops)	3.17	2.73	3.01
PERMETHRIN (for animals)	3.12	2.29	2.82
TERBUFOS	3.79	3.47	3.67
FONOFOS	0.17	0.17	0.17
TRICHLORFON	0.20	0.19	0.20
LINDANE	1.31	0.92	1.17
CARBOFURAN	1.35	1.21	1.30
CHLORPYRIFOS	8.93	7.97	8.58
MALATHION	12.78	10.00	11.76
PARATHION	0.00	0.00	0.00
CARBARYL	9.06	6.63	8.17
DIAZINON	2.91	2.42	2.73
ALDICARB	1.67	2.31	1.91
PHORATE	0.72	0.82	0.75
ALDRIN	0.00	0.00	0.00
CHLORDANE	0.05	0.00	0.03
DIELDRIN	0.00	0.00	0.00
DDT	0.00	0.00	0.00
HEPTACHLOR	0.01	0.00	0.00
TOXAPHENE	0.01	0.00	0.01
COUMAPHOS	0.44	0.28	0.38
DICHLORVOS	0.61	0.47	0.56

[a]Imputed prevalence is average of five imputations.

Use of any Pesticide

The first step in the imputation process was to impute the use of any pesticides since Phase 1 using subjects who completed both Phase 1 and 2 questionnaires. Both the enrollment and the take-home portions of the Phase 1 questionnaire were used in the modeling process. The use of any pesticides was a binary variable, and we therefore used logistic regression to model its probability based on Phase 1 responses. We considered all variables from Phase 1 that had the potential to be associated with either missingness or pesticide use (see Table 2 for candidate covariates). We first conducted a univariate analysis of Phase 1 variables, except the pesticide-specific variables. The variables most strongly predictive of use of any pesticide on the Phase 2 questionnaire were sex, marital status, farm ownership, farm size, days/year mixing pesticides, percent time personally mixing pesticides, percent time personally applying pesticides, and application of any pesticide in the prior year. Covariates associated with non-response to Phase 2 were continuous

© 2012 Nature America, Inc.

CBD v. EPA (1:21-cv-00681-CJN) ED_005427A_00025625-00002

Table 2. Phase 1 candidate covariates to predict use of any pesticide in Phase 2 (1999–2005) of AHS.

Demographics
Age (AGE_AT_ENROLLMENT)[a]
Sex (GENDER)[a]
State (SITE)[a]
County (COUNTY)
Professional/private license type (APP_TYPE)[a]
Marital status / family size (AMARITAL)[a]
Education (ASCHOOL, collapsed)[a]
Farm characteristics
Owner (AOWNFARM)[a]
Farm size (AACRES)[a]
Pesticide use
Years mixing pesticides (AYRSMIX)[a]
Days/year mixing pesticides (AMIXDPY)[a]
Percent Mix (APCTMIX)[a]
Percent Apply (APCTAPPL)[a]
Application Methods (AAPMTH1 - AAPMTH21)
Do not personally apply (AAPMTH 1)[b]
Hand spray gun application (AAPMTH 4)[b]
Backpack spray application (AAPMTH 5)[b]
In furrow or banded application (AAPMTH 8)[b]
Application Uses (APSTAP1 - APSTAP17)
Rodent control (APSTAP2)[b]
Highway right-of-way weed control (APSTAP6)[b]
Herbicide (weed killers) applications to farm crops (APSTAP9)[b]
Insecticide applications to farm animals/animal shelters (APSTAP12)[b]
Fungicides (chemicals for controlling disease on crops) (APSTAP16)[b]
Fumigants (gases or liquids that turn into gas when released) (APSTAP17)[b]
Application in past 12 mos (APSTAP18)[a]
Personal Protective Equipment (APROTEQ1- APROTEQ8)
Chemical resistant gloves (APROTEQ7)[b]
Crops and Amimals (ACRPAN1 - ACRPAN8)
No Crops or animals (ACRPAN2)[b]
Medical conditions
Diagnosis of various conditions and diseases (A_MEDCOND5 - A_MEDCOND56)
Ever diagnosed with other chronic lung disease (A_MEDCOND10)[b]
Ever diagnosed with Diabetes (A_MEDCOND16E)[b]

[a]Covariates forced into the model.
[b]Covariates selected for the final model in step-wise selection process.

age, education, state, applicator type, and years mixing chemicals.[10] These variables and covariates were forced into the logistic regression model. Other potential covariates from Phase 1 (Table 2) were included or excluded based on the SAS step-wise regression procedure, with entrance and removal criteria of $P \leq 0.001$ and $P > 0.01$, respectively. Strict criteria were set because the dataset of individuals with complete data was so large. See Table 2 for final covariates in the model.

We used the aforementioned logistic model with covariates based on Phase 1 data to compute a predicted probability of the use of any pesticides for each individual who did not complete Phase 2 ($\hat{p}_i$, $i = 1, \ldots, 20{,}968$). For the ith individual, we imputed use (yes/no) of any pesticides as follows. With $\hat{p}_i$between 0 and 1, we generated five uniform random variables between 0 and 1, Z_{ij}, $j = 1, \ldots, 5$. If $Z_{ij} \leq \hat{p}_i$, then we assigned $U_{ij} = 1$, otherwise we assigned $U_{ij} = 0$, where $U_{i1}, \ldots, U_{i5}$ were the imputed values for use of any pesticides in Phase 2.

For each individual and each imputation with an imputed "no" ($U_{ij} = 0$), the 50 pesticide-specific use variables (yes/no) and the 50 chemical-specific days/year variables in Phase 2 (Table 1) were set to zero. For each individual and each imputation with an imputed "yes" to use of any pesticide ($U_{ij} = 1$), the 50 missing chemical specific use variables and days/year were then imputed.

Use of Specific Pesticides

Using data from participants who completed both Phase 1 and 2 questionnaires, we applied the same process to generate a model for the probability of use of a specific pesticide in the interim period between Phases 1 and 2. However, we forced pesticide-specific covariates from Phase 1 (use of the specific chemical in the past year, ever mixed or applied the chemical in the past, number of years using the chemical, and days per year using the chemical) into the logistic model in addition to the 13 covariates for the model of use of any pesticide (see Table 2). The stepwise procedure in SAS identified other meaningful covariates for each pesticide, based on the entrance and removal criteria and likelihood ratio statistics. For each participant missing Phase 2 information for whom we imputed a "yes" to use of any pesticide, $U_{ij} = 1$, we generated a predicted probability for the use of a specific pesticide and randomly imputed five binary responses based on a uniform random number generator. Five responses (yes/no) were imputed for each of the 50 specific pesticides, V_{ijk} with $k = 1, \ldots, 50$. For those with Phase 1 and 2 data, it was not uncommon for participants to indicate applying or mixing of pesticides in Phase 2, while providing no affirmative response for any of 50 specific pesticides considered. This could suggest use of other pesticides or the inability to recall a specific pesticide. For that reason, we did not require that at least 1 specific pesticide be imputed as "yes", nor did we reverse the order by first imputing the 50 pesticides and then infer overall usage.

Days Per Year Use of Specific Pesticides

For each individual with an imputed "yes" to use of a specific pesticide, $V_{ijk} = 1$, we next developed a procedure to impute days/year of use. Because the Phase 2 question for days/year had an ordinal response and because data were skewed and sparse, we implemented a stratified sampling scheme using participants who completed both Phase 1 and 2 and who reported the number of days/year they used the pesticide of interest. For those missing Phase 2 data and imputed to have used a specific pesticide, we randomly selected days/year of use from the empirical frequency distribution derived from those with Phase 1 and 2 data who used the pesticide and who were in an appropriate stratum. The first step in this process was to identify an informative stratification. Table 1 indicates that the prevalence of the use of specific pesticides in Phase 2 ranged from 0% (pesticide use was discontinued) to >50%. For infrequently used pesticides, which were the majority, we could use only a limited number of Phase 1 stratification variables. By contrast, for widely used pesticides (e.g., 2,4-dichlorophenoxyacetic acid (2,4-D)), we could potentially use many stratification variables. However, to maintain consistency of methods across variables, we selected only variables most strongly associated with Phase 2 days/year use as stratification factors. After considering several possible stratification variables (age, state, applicator type, Phase 1 days use, and others; data not shown), we based the imputation of Phase 2 days/year of use of a specific pesticide on a stratification by Phase 1 days/year of use of a specific pesticide. Thus, for an applicator missing Phase 2 days/year of use of a specific pesticide, we identified the Phase 1 days/year of use category, then randomly sampled (with replacement) a value from the frequency distribution for Phase 2 days/year of use that corresponded to the same Phase 1 days/year of use category.

Finally, for those missing Phase 2 data, we also needed to impute the most recent year of farming activity. This year (see questions 10 and 13 of the private and commercial Phase 2 Questionnaires,[7] respectively at www.aghealth.org/questionnaires.html) was critical for calculating cumulative exposure to pesticides. Because reference year is an integer with a 12-year range (1993–2004), we again employed stratified sampling with replacement. The primary stratification variable was the use of any pesticide in Phase 2. If the imputed value for use of any pesticide was "no", then we defined 10 strata (applicator type [commercial or private] by enrollment year [1993–1997]). If the imputed value for use of any pesticide was "yes", then we defined 50 strata (applicator type by enrollment year by age at AHS enrollment in quintiles). For each stratum, we computed the frequency distribution of the most recent year of farming activity from those with complete Phase 1 and 2 data. We constrained the imputed reference year to occur after the enrollment year and, when an individual

© 2012 Nature America, Inc.

CBD v. EPA (1:21-cv-00681-CJN) ED_005427A_00025625-00003

was known to have died, before the year of death. If the enrollment year was equal to or within 1 year of death, we set the reference year to the enrollment year.

RESULTS

Imputation Assessment

We assessed the imputation method by holding out a randomly selected subset (20%, $n=7269$) of the observed complete data and imputing multiple values for Phase 2 as though the data were missing. The "true" use of any pesticides in this subset was 85.68% with standard error 0.41%. The average of the five imputations indicated a prevalence of 85.25% with imputation adjusted standard error of 0.59%. This indicates that the logistic regression model underpinning the multiple imputation procedure did indeed preserve essential features of the data. Recall, the modeling process we used first generated a probability of use (the use of any pesticide, or the use of a specific pesticide) for each individual, $\hat{p}_i$. To assess the accuracy of the implemented prediction model, and how it compares with a "naïve" reference prediction (e.g., change prediction based on observed prevalence), we calculated the Brier[11] and Brier skill scores,[12] commonly utilized in atmospheric probability forecasting and risk prediction modeling. In the holdout set, let X_i be the observed use of any pesticides, $X_i=0$ or 1, $i=1,\ldots,n$, for the ith individual in the holdout data. Let $\hat{p}_i$ be the predicted probability of use from the logistic model. The Brier score estimator is $B=1/n\times\sum_{i=1}^{n}(X_i-\hat{p}_i)^2$ and is equivalent to the mean squared error of prediction; the smaller the value the better the prediction. To assess the utility of any prediction model, it can be compared to a naïve prediction using the skill score, $SS=1-B/B_{Rf}$, where B_{Rf} is the Brier score estimator using a reference, or naïve forecast, p' in place of the model $\hat{p}_i$ prediction. In this evaluation, we use the observed Phase 2 prevalence of pesticide use in the complete data ($N=36{,}342$) less the holdout observations ($n=7269$) as the reference prediction, $p'=1/n'\times\sum_{i=1}^{n'}X_i$, where $n'=N-n$. For use of any chemicals, $B=0.1092$, $B_{Rf}=0.1227$, for a $SS=0.1103$, an 11% improvement in accuracy using the predictive model over simple prediction based on observed Phase 2 usage. Parker and Davis[13] proposed a similar metric to the skill score, which was the sum of sensitivity and specificity, whereby the sum must be >1 for the observed accuracy to be larger than chance. Figure 1 is a plot of Brier skill score *versus* the sum of sensitivity and specificity (pooling all five imputations for calculations) for overall pesticide use and commonly used pesticides (percent usage $>0.05\%$). The two metrics are highly correlated ($r=0.925$) and essentially measure the same thing, proportional improvement of prediction model over naïve/chance prediction.

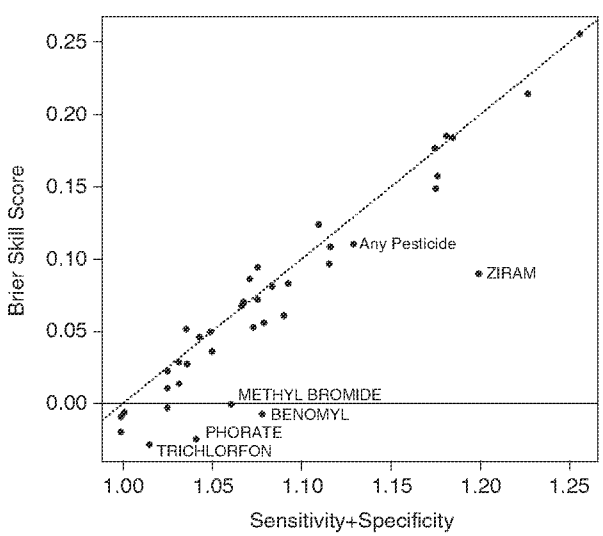

Figure 1. Scatterplot of Brier skill score *versus* sensitivity + specificity for commonly used pesticides ($P>0.05\%$).

Use of Specific Pesticides

Table 3 gives the observed ("true") and imputed prevalence for the 38 pesticides where observed prevalence $>0.05\%$. The mean and standard error of a variable that includes multiply imputed values is well known.[1] Therefore, for any chemical, let X_i be the observed use of the pesticide of interest, $X_i=0$ or 1, $i=1,\ldots,n$ for the ith individual in the holdout data. The estimated mean and variance of the percent usage (prevalence) in the holdout data is: $p=(1/n)\times\sum_{i=1}^{n}X_i$ and $s^2=p\times(1-p)/n$, respectively. It follows that the usual standard error of the estimated prevalence p, is s. The prevalence from one of the m multiply imputed datasets is $\tilde{p}_j=(1/n)\times\sum_{i=1}^{n}\tilde{X}_{ij}$ where $\tilde{X}_{ij}=0$ or 1, the imputed use of the pesticide of interest for individual i. Then, the overall prevalence estimate and its variance from the m (in this case 5) imputed datasets are $\tilde{p}=(1/m)\times\sum_{j=1}^{m}\tilde{p}_j$ and $\tilde{s}^2=1^{m}(\tilde{p}_j-\tilde{p})^2$, where $\tilde{s}_j^2=(1/n)\times\tilde{p}_j\times(1-\tilde{p}_j)$ and $\tilde{s}$ is the standard error of $\tilde{p}$.

As expected, the multiple imputation estimates of the standard error are slightly higher than the "true" standard error because the variability of the random imputations are included in the estimates, and pesticides with the highest prevalence (e.g., atrazine, 31.47%) have the largest standard errors while rarely used pesticides (e.g., methyl bromide, 0.41%) have little variability. Imputed prevalence is generally lower than observed both in Table 1 (across Phase 2 responders and non-responders) and Table 3 (the validation set). The Brier skill scores in Table 3 show a range of improvement from none to 25% over the naïve, or reference prediction model. Models for aldicarb and chlorothalonil appear to perform the best (*SS* of 0.256 and 0.214, respectively), while the majority of pesticides fall between $SS=0.05$ and 0.20, including 2,4-D and atrazine with an 18% improvement in accuracy over naïve predictions. Some of the least prevalent pesticides did not benefit much from the implemented modeling scheme, and some of their skill scores were slightly negative (e.g., EPTC, phorate, benomyl, fonofos, and trichlorphon). The variability corresponding to rare event predictions can be large relative to the naïve estimates, and can yield negative skill scores. Skill scores close to zero (negative or positive) indicate that the predictive model was of limited additional value for these pesticides.

Figure 2 is a plot of the relative errors of the imputed prevalence estimate, $\tilde{p}$ to their respective true estimate, p, i.e., $\varepsilon=(\tilde{p}-p)/p$, for the 38 pesticides with $>0.05\%$ use. Relative errors, ε, are centered about zero, and mostly fall within ±0.20. For only a few of the rare pesticides ($<1.0\%$ usage) used in Phase 2 does the imputed prevalence differ from the "true" prevalence by more than 20% (e.g., petroleum oil/petroleum distillates, methyl bromide, maneb/mancozeb, trichlorfon, metalaxyl, dichlorvos, coumaphos, and phorate).

Days Per Year Use of Specific Pesticides

We imputed days per year for a specific pesticide by sampling with replacement from the observed Phase 2 data stratified by Phase 1 days use of that pesticide. Figure 3 shows the box plots of the observed data from the validation dataset alongside the imputed data for days/year for three pesticides. Alachlor, diazinon, and 2,4-D were chosen for illustration because they were widely used and represent common usage patterns in the AHS cohort. The distributions of the imputed values for the three pesticides were very similar to those of the "true" data. The means (solid

© 2012 Nature America, Inc.

CBD v. EPA (1:21-cv-00681-CJN) ED_005427A_00025625-00004

Table 3. Prevalence, standard error and Brier scores of pesticide use in holdout dataset ($N = 7269$) of the AHS.

Pesticide name	*Observed*		*Imputed*[a]		*Reference Brier*	*Brier score*	*Brier skill score*
	Prevalence (%)	*Standard error*	*Prevalence (%)*	*Standard error*			
METHYL BROMIDE	0.43	0.08	0.56	0.12	0.004	0.004	−0.001
ALUMINUM PHOSPHIDE	0.59	0.09	0.71	0.13	0.006	0.005	0.149
BENOMYL	0.37	0.07	0.29	0.08	0.004	0.004	−0.007
CHLOROTHALONIL	2.39	0.18	2.33	0.26	0.023	0.018	0.214
CAPTAN	2.12	0.17	2.11	0.28	0.021	0.020	0.053
MANEB/MANCOZEB	0.15	0.05	0.18	0.06	0.002	0.002	−0.020
METALAXYL	2.66	0.19	2.09	0.23	0.026	0.023	0.124
ZIRAM	0.12	0.04	0.11	0.05	0.001	0.001	0.090
ATRAZINE	31.85	0.55	27.64	0.69	0.217	0.177	0.185
DICAMBA	19.16	0.46	17.39	0.48	0.155	0.128	0.177
CYANAZINE	1.75	0.15	1.50	0.21	0.017	0.017	0.029
CHLORIMURON-ETHYL	2.93	0.20	2.93	0.36	0.028	0.027	0.050
METOLACHLOR	14.87	0.42	13.23	0.55	0.127	0.113	0.109
EPTC	0.30	0.06	0.30	0.09	0.003	0.003	−0.003
ALACHLOR	2.82	0.19	2.43	0.32	0.027	0.026	0.052
METRIBUZIN	2.19	0.17	1.75	0.22	0.021	0.021	0.022
PARAQUAT	1.91	0.16	1.88	0.22	0.019	0.017	0.086
PETRO. OIL/PETRO. DISTILLATES	0.47	0.08	0.60	0.13	0.005	0.005	−0.006
PENDIMETHALIN	11.24	0.37	10.36	0.48	0.100	0.093	0.068
IMAZETHAPYR	7.76	0.31	7.36	0.39	0.072	0.067	0.070
GLYPHOSATE	52.73	0.59	45.42	0.83	0.249	0.225	0.097
TRIFLURALIN	10.58	0.36	10.21	0.58	0.095	0.080	0.157
2,4-D	36.92	0.57	33.30	0.86	0.233	0.190	0.184
PERMETHRIN (for crops)	3.36	0.21	2.71	0.24	0.032	0.031	0.036
PERMETHRIN (for animals)	3.05	0.20	2.83	0.33	0.030	0.028	0.061
TERBUFOS	3.80	0.22	3.38	0.33	0.037	0.033	0.095
FONOFOS	0.17	0.05	0.15	0.07	0.002	0.002	−0.009
TRICHLORFON	0.17	0.05	0.13	0.05	0.002	0.002	−0.028
LINDANE	1.39	0.14	1.07	0.18	0.014	0.013	0.046
CARBOFURAN	1.36	0.14	1.14	0.24	0.013	0.013	0.014
CHLORPYRIFOS	8.87	0.33	7.90	0.46	0.081	0.074	0.081
MALATHION	12.88	0.39	11.50	0.49	0.112	0.103	0.083
CARBARYL	9.34	0.34	7.69	0.65	0.085	0.079	0.072
DIAZINON	2.94	0.20	2.71	0.28	0.029	0.028	0.027
ALDICARB	1.66	0.15	1.57	0.18	0.016	0.012	0.256
PHORATE	0.59	0.09	0.69	0.17	0.006	0.006	−0.024
COUMAPHOS	0.56	0.09	0.33	0.10	0.006	0.005	0.056
DICHLORVOS	0.65	0.09	0.48	0.12	0.006	0.006	0.010

[a]Imputed prevalence is average of five imputations and standard error is calculated via equation in text.

squares) were more sensitive to outliers for the less frequently used pesticides since fewer than 200 individuals reported use of those pesticides in the 20% holdout set. Comparing the observed reference year with its imputed value, Figure 4 indicates that for 90% of participants with reference year 1998 through 2004, the imputed years were centered around the expected year. When the "true" reference year is 1994–1997 the sampled imputation values were higher than expected and indicated bimodality. This was due to the ordinal nature of reference year and the scheduled pattern of interviews. The first interviews were conducted between 1993 and 1997 (Phase 1), while the follow-up Phase 2 interviews occurred between 1999 and 2005. When an individual participated in Phase 2, the most likely responses for reference year were 1) the year prior to the Phase 2 interview, 2) 5 years prior (year of Phase 1), or 3) the last year of farming prior to enrollment. This bimodal behavior seen in approximately 10% of the holdout dataset tended to occur in individuals who reported "no farming" or "no pesticide application" in Phase 2, and therefore a reference year for pesticide use in Phase 2 was irrelevant.

Post-assessment of the holdout dataset, all of the observed data were used to generate the complete predictive model and populate the sampling data. The final multiple imputations were generated and prevalence estimates for the 50 pesticides in the imputed subset and overall are shown in Table 1.

DISCUSSION

The lifetime exposure of an individual to a specific pesticide or set of pesticides is the primary quantity of interest in the AHS for studying the association between exposure and disease outcomes. A substantial number of AHS participants were non-responders to a Phase 2 questionnaire used to update lifetime pesticide use following enrollment. In analyses, imputation is generally preferable to omitting individuals who did not complete Phase 2 (in our case, 37% of enrolled individuals) due to possible selection bias in the subset with complete data and decreased precision of parameter estimates using only a subset of the data. This paper illustrates the use of a multi-step, conditional imputation procedure combining parametric modeling and sampling from an empirical distribution for several variable types. Using multiple imputation, the variables necessary to calculate exposure for those missing Phase 2 data are replaced by five imputed values. For validation purposes, we estimated prevalence of pesticide use and showed the form of the variance estimate for prevalence resulting from multiple imputation. Prevalence estimates for the Phase 2 non-responders were slightly lower than in the responders and this is likely due to the slightly different makeup of individuals in each. Logistic regression is known to perform sub-optimally when modeling rare events,[14] which may

© 2012 Nature America, Inc.

CBD v. EPA (1:21-cv-00681-CJN) ED_005427A_00025625-00005

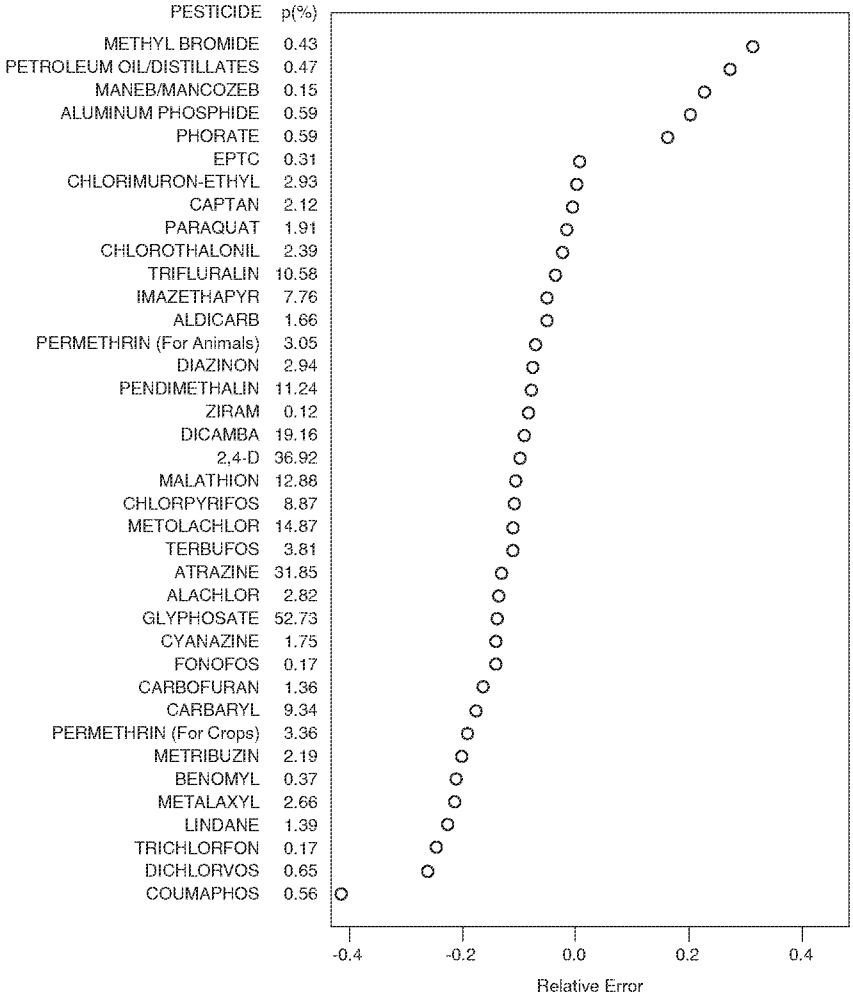

Figure 2. Relative errors of imputed prevalence or percent usage (p) for commonly used pesticides ($P > 0.05\%$).

explain the low imputed prevalence estimates in the validation set; the underestimation makes our imputation slightly conservative, favoring specificity over sensitivity.

Rubin's method of scalar estimands in multiple imputation procedures[15] is generalizable and can be used to calculate standard errors and confidence intervals for any estimator including risk ratios, absolute risk, and hazard ratios. We applied fractional hot deck imputation[16] to impute days/year use of a pesticide, for which other variance estimators have been proposed;[16–19] however, their utility has not been explored here.

Multiple imputation, in contrast to single imputation, accounts for the uncertainty of predicting missing data with limited loss of efficiency (nearly 94% efficient when imputed five times with 35% missing data, as opposed to 74% efficiency with a single imputation[1]). The observed data, together with the five imputed values for missing variables, generate five complete datasets to be analyzed by standard statistical techniques resulting in five slightly different results. These results and their variance/covariance matrices are combined to represent the variability induced by the imputing process. For simplicity, modeling and sampling were performed using the single set of observed complete data, as opposed to first bootstrapping the complete data to perform a proper imputation, which accounts for variability of regression parameter estimates used in the imputation.[1] An assessment of proper *versus* improper imputation on a dataset similar to the AHS shows mixed results.[20] Multiple imputation was chosen for pesticide use in the AHS over other approaches such as probability weighting or the EM algorithm[21] because of its familiarity and ease of use. Providing a single set of multiply imputed data will facilitate consistent results in future analyses.

A key assumption of any imputation is that missingness is independent of the unobserved outcome of interest or unobservable confounders (i.e., missing at random). The reduction of bias and increase in precision from multiple imputations is dependent on the covariates associated with both non-response and the endpoint variable,[22] and factors associated with non-participation, which were included in our imputation model. For our imputation analysis, the "outcome" of interest is the missing pesticide use itself; Montgomery et al.[10] show there is little evidence for selection bias in Phase 2 of the AHS, however missing at random is an untestable assumption without additional data; thus it is possible that non-responders differ from responders in variables we have not measured. It is worth emphasizing that the set of individuals with both Phase 1 and 2 responses had a full range of exposure, including those who were no longer farming, and therefore our data-driven imputation approach did not necessitate that non-responders be imputed as active pesticide users. To implement multiple imputation, missingness may be conditional on observable covariates from Phase 1 and our models incorporated covariates associated with Phase 2 pesticide use in constructing the values for missing data.

© 2012 Nature America, Inc.

CBD v. EPA (1:21-cv-00681-CJN) ED_005427A_00025625-00006

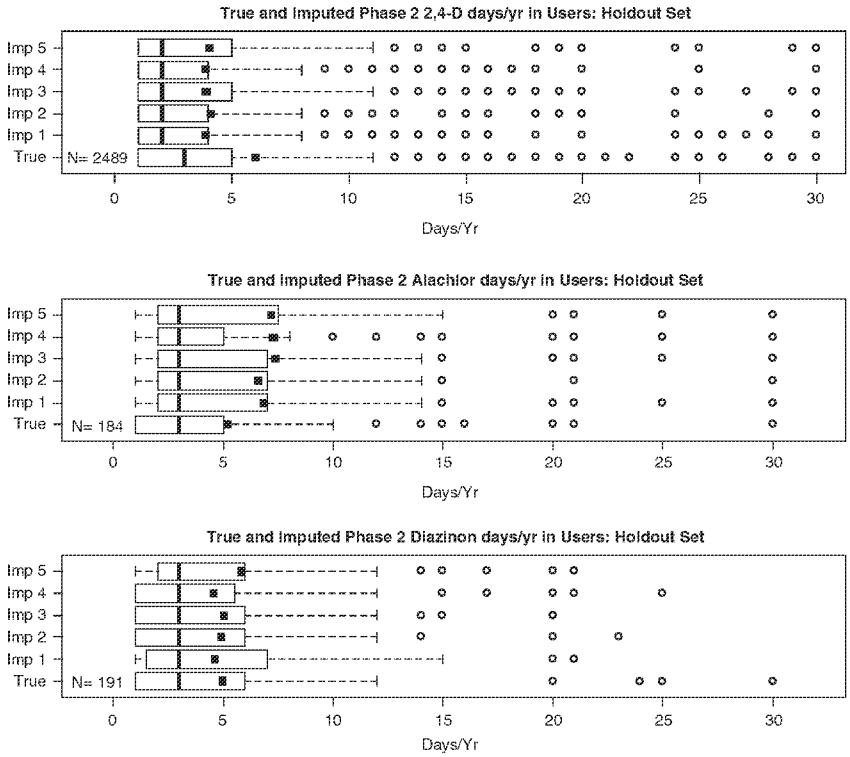

Figure 3. Box plots of observed and imputed days/year use of 2,4-D, alachlor, and diazinon in the holdout subset of the AHS.

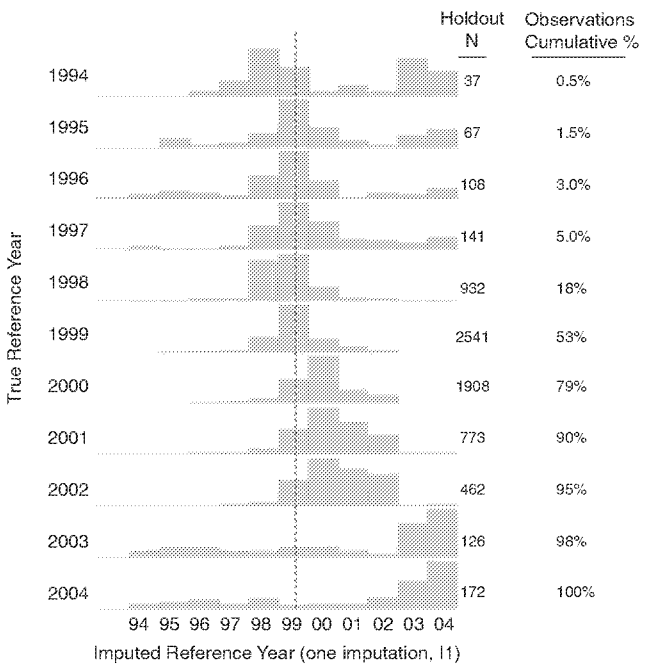

Figure 4. Histogram display of the distribution of imputed Phase 2 reference year by true, observed reference year in the holdout dataset of the AHS.

As was done for information collected from participants who completed the Phase 2 questionnaire, for epidemiologic analyses, the imputed pesticide use information has been cumulatively added to information collected in Phase 1. This multiple imputation will allow for bias reduction and improved efficiency in future analyses of the AHS.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work is supported by the Intramural Research Program of the National Cancer Institute at the National Institutes of Health (grant number Z01-CP010119); and the National Institute of Environmental Health Sciences at the National Institutes of Health (grant number Z01-ES049030). The United States Environmental Protection Agency through its Office of Research and Development collaborated in the research described here. It has been subjected to Agency review and approved for publication. The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the National Institute for Occupational Safety and Health.

REFERENCES

1 Rubin D.B. *Multiple Imputation of Nonresponse in Surveys*. J.Wiley and Sons: New York, NY, 1987.
2 Schafer J.L., Ezzatti-Rice T.M., Johnson W., Khare M., Little R.J.A., and Rubin D.B. The NHANES III multiple imputation project. *Proc Survey Res Methods Section Am Stat Assoc* 1996, 28–37.
3 Mislevy R.J., Johnson E.G., and Muraki E. Scaling procedures in NAEP. *J Educational Stat* 1992: **17**: 131–154.
4 Stuart E.A., Azur M., Frangakis C., and Leaf P. Multiple imputation with large data sets: a case study of the children's mental health initiative. *Am J Epidemiol* 2009: **169**(9): 1133–1139.
5 Kang T., Kraft P., Gauderman W.J., and Thomas D. Multiple imputation methods for longitudinal blood pressure measurements from the Framingham Heart Study. *BMC Genet* 2003: **4**(suppl 1): S43.
6 Alavanja M.C., Sandler D.P., McMaster S.B., Zahm S.H., McDonnell C.J., and Lynch C.F., et al. The Agricultural Health Study. *Environ Health Perspect* 1996: **104**: 362–369.
7 National Cancer Institute, National Institutes of Health. Agricultural Health Study (AHSQ). Full Text of Questionnaires. 2010. (www.aghealth.org/questionnaires.html). (Accessed November 8, 2010).
8 Tarone R.E., Alavanja M.C., Zahm S.H., Lubin J.H., Sandler D.P., and McMaster S.B., et al. The Agricultural Health Study: factors affecting completion and return of self-administered questionnaires in a large prospective cohort study of pesticide applicators. *Am J Ind Med* 1997: **31**: 223–242.

© 2012 Nature America, Inc.

CBD v. EPA (1:21-cv-00681-CJN) ED_005427A_00025625-00007

9 Dosemeci M., Alavanja M.C., Rowland A.S., Mage D., Zahm S.H., and Rothman N., et al. A quantitative approach for estimating exposure to pesticides in the agricultural health study. *Ann Occup Hyg* 2002: **46**(2): 245–260.
10 Montgomery M.P., Kamel F., Hoppin J.A., Beane Freeman L.E., Alavanja M.C., and Sandler D.P. Effects of self-reported health conditions and pesticide exposures on probability of follow-up in a prospective cohort study. *Am J Ind Med* 2010: **53**: 486–496.
11 Brier G.W. Verification of forecasts expressed in terms of probability. *Monthly Weather Rev* 1950: **78**(1): 1–3.
12 Murphy S.H. Hedging and skill scores for probability forecasts. *J Appl Meteor* 1973: **12**(1): 215–223.
13 Parker R.A., and Davis R.B. Evaluating whether a binary decision rule operates better than chance. *Biom J* 1999: **41**: 25–31.
14 King G., and Zeng L. Logistic regression in rare events data. *Political Anal* 2001: **9**: 137–163.
15 Little R.J.A., and Rubin D.B. *Statistical Analysis with Missing Data*, 2nd edn J.Wiley and Sons: New York, NY, 2002.
16 Kim J.K., and Fuller W.A. Fractional hot deck imputation. *Biometrics* 2004: **91**(3): 559–578.
17 Rao J.N.K., and Shao J. Jackknife variance estimation with survey data under hot deck imputation. *Biometrika* 1992: **79**: 811–822.
18 Rubin D.B., and Schenker N. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *J Am Stat Assoc* 1986: **81**: 366–374.
19 Tollefson M., and Fuller W.A. Variance estimation for samples with random imputation. *American Statistical Association Proceedings of the Section of Survey Research Methods* 1992: **15**: 758–763.
20 Heitjan D.F., and Little R.J.A. Multiple imputation for the fatal accident reporting system. *Appl Stat* 1991: **40**: 13–29.
21 Dempster A.P., Laird N.M., and Rubin D.B. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B* 1977: **39**(1): 1–38.
22 Spratt M., Carpenter J., Sterne J.A., Carlin J.B., Heron J., and Henderson J., et al. Strategies for multiple imputation in longitudinal studies. *Am J Epidemiol* 2010: **172**(4): 478–487.

© 2012 Nature America, Inc.

CBD v. EPA (1:21-cv-00681-CJN) ED_005427A_00025625-00008