American Journal of Epidemiology
© The Author(s) 2017. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

Vol. 186, No. 4
DOI: 10.1093/aje/kwx098
Advance Access publication:
July 7, 2017

Practice of Epidemiology

Assessing the Potential for Bias From Nonresponse to a Study Follow-up Interview: An Example From the Agricultural Health Study

Jessica L. Rinsky, David B. Richardson, Steve Wing, John D. Beard, Michael Alavanja, Laura E. Beane Freeman, Honglei Chen, Paul K. Henneberger, Freya Kamel, Dale P. Sandler, and Jane A. Hoppin*

* Correspondence to Dr. Jane A. Hoppin, Department of Biological Sciences, Campus Box 7633, North Carolina State University, Raleigh, NC 27695-7633 (e-mail: jahoppin@ncsu.edu).

Initially submitted October 11, 2015; accepted for publication September 29, 2016.

Prospective cohort studies are important tools for identifying causes of disease. However, these studies are susceptible to attrition. When information collected after enrollment is through interview or exam, attrition leads to missing information for nonrespondents. The Agricultural Health Study enrolled 52,394 farmers in 1993–1997 and collected additional information during subsequent interviews. Forty-six percent of enrolled farmers responded to the 2005–2010 interview; 7% of farmers died prior to the interview. We examined whether response was related to attributes measured at enrollment. To characterize potential bias from attrition, we evaluated differences in associations between smoking and incidence of 3 cancer types between the enrolled cohort and the subcohort of 2005–2010 respondents, using cancer registry information. In the subcohort we evaluated the ability of inverse probability weighting (IPW) to reduce bias. Response was related to age, state, race/ethnicity, education, marital status, smoking, and alcohol consumption. When exposure and outcome were associated and case response was differential by exposure, some bias was observed; IPW conditional on exposure and covariates failed to correct estimates. When response was nondifferential, subcohort and full-cohort estimates were similar, making IPW unnecessary. This example provides a demonstration of investigating the influence of attrition in cohort studies using information that has been self-reported after enrollment.

attrition; epidemiologic methods; inverse probability weights; loss to follow-up; occupational/environmental epidemiology; prospective studies; selection bias

Abbreviations: AHS, Agricultural Health Study; IPEW, inverse probability of exposure weight; IPSW, inverse probability of selection weight.

Prospective cohort studies are important tools for identifying causes of disease (1–3). However, studies that follow participants over time are susceptible to attrition, including loss to follow-up and death (3–7). Attrition in cohort studies that collect information through interviews or exams conducted after enrollment often leads to missing exposure or outcome information. Analyses limited to interview respondents may provide estimates of disease occurrence or exposure-disease associations that differ from estimates that would have been obtained from the source population (8), indicating bias (3, 6, 9). As previously demonstrated through simulations (4, 6, 10) and directed acyclic graph theory (9, 11, 12), bias can occur if attrition is related to both the exposure and outcome under study or to factors related to both the exposure and outcome. Therefore, identification of factors related to attrition can help evaluate potential for bias in studies of outcomes self-reported after enrollment.

Identifying factors associated with attrition can be useful in determining whether selection bias might affect results. In addition, with attrition-related factors identified, investigators can determine the utility of analytical methods (e.g., inverse probability weighting, multiple imputation of missing data,

Downloaded from https://academic.oup.com/aje/article/186/4/395/3806857 by Centers for Disease Control and Prevention user on 19 December 2023

sensitivity analyses) to illustrate or mitigate the influence of attrition. As these analytical methods have become easier to implement, analyses of cohort studies have begun to include formal evaluation of the potential influence of attrition and the use of such methods to mitigate potential bias (13–15).

The Agricultural Health Study (AHS) is a longitudinal cohort study that enrolled 52,394 private pesticide applicators, hereafter "farmers," who applied for restricted-use pesticide licenses in Iowa and North Carolina between 1993 and 1997 (16, 17). The AHS was designed to evaluate the associations between farming-related exposures and health outcomes among farmers and their spouses (16). Cancer incidence, mortality, and end-stage renal disease data are obtained through linkage with state and federal sources, and therefore analyses of exposures measured at enrollment and these outcomes are unaffected by attrition. Information about current farming activities and other morbidity is collected via interviews occurring approximately every 5 years. Forty-six percent of farmers enrolled in the AHS responded to the 2005–2010 interview, creating the potential for attrition to affect distributions of exposures or outcomes self-reported at that interview.

We sought to identify characteristics associated with response to the 2005–2010 interview, evaluate the extent to which attrition might influence results of analyses restricted to respondents, and consider the utility of inverse probability weights to correct for potential selection bias in studies of incident outcomes reported only by 2005–2010 interview respondents. The goal of this analysis was to provide information useful for cohort studies such as the AHS that rely on self-reported information on outcomes collected after enrollment.

METHODS

To enroll in the AHS, farmers provided information on demographic factors and lifestyle, medical history, and farming activities via questionnaire. Investigators attempted to recontact farmers between 1999 and 2003 using a computer-assisted telephone interview to update information on farming activities. Investigators have previously examined factors associated with nonresponse to this interview (18).

Another interview was conducted between 2005 and 2010. At this time, farmers were asked to provide updated information on lifestyle and farming activities and to report several medical conditions. Reasons for nonresponse to this interview have been described previously (19). Briefly, 3 groups of farmers were not contacted for this interview: 1) farmers who refused further contact prior to the interview, 2) farmers who died prior to the interview, and 3) farmers who were excluded for a variety of administrative reasons or because they had not participated in any study activity since enrollment (19). Farmers who were invited to complete the computer-assisted telephone interview did so or they refused, could not be reached, or were too ill to respond. For the present analysis, farmers who completed the interview are referred to as "respondents." "Nonrespondents" include farmers who were not invited to complete the interview and those who were invited but declined or could not be reached. For this analysis, those who died prior to the 2005–2010 interview were included as nonrespondents because mortality is one mechanism by which disease outcomes may result in missing self-reported outcome information after enrollment.

This evaluation was approved by all relevant institutional review boards. Participants indicated initial informed consent by completing the enrollment questionnaire. Questionnaires are available on the study Web site (20). This analysis used AHS data release P1REL0907.00 and REL0905.00.

Associations between enrollment information and response to the 2005–2010 interview

We considered several attributes reported at enrollment, and commonly used in other AHS analyses, as predictors of response to the 2005–2010 interview. These variables included demographic and lifestyle factors (age, state, sex, race/ethnicity, education, marital status, smoking status, alcohol consumption), medical conditions (heart disease, asthma, other chronic lung disease, kidney disease, diabetes, Parkinson disease, depression, tuberculosis, pneumonia), personal use of pesticides (ever use, percentage of time using, lifetime use, days per year of use, ever use of functional groups and chemical classes), and other farm characteristics (farm size, work in hog and poultry confinement, number of livestock and poultry, major income-producing animals).

We examined the distribution of each variable by response status (respondent vs. nonrespondent) and by reason for nonresponse (death vs. refusal/exclusion). We estimated crude associations between variables and response and categorized variables to preserve the shape of the association. We then used logistic regression models and a backward elimination approach to identify a set of variables that described the relationship between covariates reported on the enrollment questionnaire and response to the 2005–2010 interview. First, we removed variables with more than 10% missing data from consideration. Then we removed variables with a nonsignificant χ^2 statistic (2-sided test; $\alpha = 0.05$). Demographic and lifestyle variables were hypothesized a priori to be the strongest predictors of response and were the last variables removed from the model. To preserve precision, we also removed variables with <1% of the population reporting the factor. Finally, we removed variables that were not strongly associated with response ($-0.35 < \beta < 0.35$).

We report regression coefficients (β) and standard errors to show the direction, magnitude, and precision of the association between each variable and response. Wald χ^2 values are reported to indicate the contribution of the variable to the prediction of response. The presented results indicate the set of enrollment variables that are predictive of and strongly associated with response.

Assessing selection bias

In the AHS, disease occurrence was reported by respondents during follow-up interviews. For this reason, case status is unavailable for nonrespondents. In contrast, cancer incidence was ascertained by linkage of the enrollment cohort with state (Iowa, North Carolina) cancer registries through December 31, 2010. Because complete-case ascertainment for cancer is available for all farmers (except for a small proportion who

Downloaded from https://academic.oup.com/aje/article/186/4/395/3806857 by Centers for Disease Control and Prevention user on 19 December 2023

left North Carolina or Iowa), we compared cancer incidence for 3 cancer outcomes to identify situations in which restricting analyses to 2005–2010 respondents might result in biased effect estimates. These examples are meant to aid in understanding the potential influence of attrition on the estimation of associations between agricultural exposures and self-reported morbidity when analyses are restricted to respondents of the 2005–2010 interview.

For these examples, we defined 2 cohorts within the AHS. The full cohort included farmers who enrolled in the AHS with complete information on relevant covariates ($n = 47{,}007$). The second cohort, a subset of the full cohort ("subcohort" hereafter) includes farmers who responded to the 2005–2010 interview with complete information on relevant covariates ($n = 22{,}214$). The subcohort includes study subjects that would be included in a complete case-analysis. Within both cohorts, we examined associations between ever smoking and 1) lung cancer, a strong, well-established association (21); 2) bladder cancer, a weaker, well-established association (22); and 3) prostate cancer, an association usually observed to be null (23, 24). These outcomes differ in mortality and disability rates that may influence a person's ability to respond to an interview. We assigned smoking status (ever vs. never) based on enrollment information. Web Figure 1 (available at https://academic.oup.com/aje) illustrates the relationships between ever smoking, each incident cancer outcome, a vector of covariates, and selection for each association.

For each smoking-cancer association, we used inverse probability of exposure weights (IPEWs) to address confounding. The application of IPEWs is a form of direct standardization that creates a "pseudopopulation" in which the distributions of confounding variables are similar across exposure groups (25–27). Based on the literature, we chose the following confounders and nonconfounding risk factors to estimate weights for each cancer outcome (28): age at censoring, state, sex, education, race/ethnicity, and marital status. Alcohol consumption was included for the smoking-bladder cancer association. The association for ever smoking and prostate cancer was restricted to male farmers. To derive exposure weights, we used logistic regression models to estimate the predicted probability of ever smoking, conditional on covariates. Next, we assigned each individual a weight equal to the inverse of the predicted probability that the individual had the observed smoking status. To stabilize the weights, we multiplied each weight by the marginal probability of the individual's observed smoking status.

We applied IPEWs to log- and linear-binomial models to estimate standardized cumulative incidence, risk differences, and risk ratios for the 3 associations. We opted to use IPEWs to facilitate estimation of risk ratios and differences using log and linear binomial models, which often fail to converge when adjusting for multiple factors. To account for within-subject correlation induced by weighting, we used robust variance estimates to estimate standard errors and 95% confidence intervals (26). We considered the estimated association for the full cohort as the target parameter of interest.

In these examples, we examine incidence proportions so results will apply to many outcomes reported as part of the 2005–2010 interview with limited information on timing of onset/diagnosis.

Inverse probability weighting for selection bias

A second set of weights—inverse probability of selection weights (IPSWs)—was estimated to address nonresponse to the 2005–2010 interview. We estimated stabilized IPSWs using logistic regression models. The numerator for each individual's stabilized IPSW was the marginal probability of the observed response status in the overall study population. The denominators were calculated in 2 ways: First, the denominator was equal to each individual's predicted probability of the observed response status conditional on smoking, the vector of variables identified as predictors of response, and disease status. These weights, referred to as IPSW|E,**Z**,D—ever smoking (E), a vector of covariates (**Z**), and each incident cancer outcome (D)—are necessary to remove selection bias when attrition is related to exposure, covariates, and outcome. In practice, nonrespondents are missing disease status, and simpler weights, conditional on E and **Z**, are often used. Simpler weights may suffice when attrition is related to exposure and covariates but not the outcome. Therefore, we estimated a second set of weights with denominators for each individual equal to the predicted probability of his or her response status conditional only on E and **Z**. We refer to this set of weights as IPSW|E,**Z**.

Among the subcohort, using log- and linear-binomial models, we estimated cumulative incidence, risk ratios, and risk differences for the 3 associations after applying IPSW|E,**Z**,D and IPSW|E,**Z**. For each analysis, adjustment for confounding and selection was achieved by applying a product of the IPEW and IPSW (25, 27, 29). Weights were well-behaved, with means close to 1 and no extreme values (<0.05 or >20). Overall results were robust to weight truncation (25), so untruncated weights were used (Web Table 1). Sensitivity analyses were conducted using separate weights to represent those who died and nonrespondents who remained alive, and sensitivity analyses were conducted using weights estimated just for living nonrespondents (excluding nonrespondents who died prior to the 2005–2010 interview). Results of these sensitivity analyses were similar to the results presented in the main text and are shown in Web Tables 2–4.

All analyses were performed using SAS, version 9.3 (SAS Institute, Inc., Cary, North Carolina).

RESULTS

In total, 24,171 farmers responded to the 2005–2010 interview, and 28,223 farmers did not (Table 1). Nonrespondents included 3,541 farmers who died prior to interview.

Associations between enrollment information and response to the 2005–2010 interview

Age, state, race/ethnicity, education, marital status, smoking status, and alcohol consumption were strongly associated with response to the 2005–2010 interview (Table 2). Response was smaller in proportion among enrollees aged <40 or ≥70 years than among those aged 40–49 years. The proportion of nonresponse due to death increased with age at enrollment. Enrollment in North Carolina, race/ethnicity other than non-Hispanic white,

Downloaded from https://academic.oup.com/aje/article/186/4/395/3806857 by Centers for Disease Control and Prevention user on 19 December 2023

Table 1. Response Status of Farmers Enrolled in 1993–1997 at the 2005–2010 Interview, Agricultural Health Study, Iowa and North Carolina

Response Status	No. of Farmers	%
Respondents	24,171	46[a]
Nonrespondents	28,223	54
Deceased	3,541	7
Refused future contact before the interview	849	2
2005–2010 refusal	5,719	11
Exclusions[b]	7,871	15
Unable to contact	9,810	19
Too ill to participate	433	1
Total enrolled (1993–1997)	52,394	100

[a] Respondents in 2005–2010 made up 46% of originally enrolled farmers; 60% of the farmers had been contacted for the 2005–2010 interview.

[b] Excluded farmers were those who did not respond to any AHS activities after enrollment (n = 7,397) and those who participated in pilot interviews or related activities (n = 474).

having less than a high school diploma, not being married/living as married, and heavy drinking were associated with a decrease in response. Current smokers at enrollment, regardless of the number of pack-years, responded less frequently than never smokers. The proportion of nonrespondents due to death increased with pack-years of smoking.

Farmers reporting a doctor diagnosis of heart disease, diabetes, or Parkinson disease at enrollment responded less frequently than their counterparts. In addition, those indicating personal pesticide use or raising animals responded more frequently than their counterparts. However, we did not include variables indicating medical conditions, pesticide use, and farm characteristics measured at enrollment because none met the a priori inclusion criteria (results not shown).

Assessing selection bias

A greater proportion of ever smokers were nonrespondents compared with never smokers (Table 3). Incident lung cancer had a strong, inverse association with response; incident bladder cancer was not associated with response; and incident prostate cancer had a weak, positive association with response. The proportion of nonresponse due to mortality was greatest for lung cancer, followed by bladder cancer and then prostate cancer.

The number and proportion of respondents by smoking and cancer outcome is shown in Table 4. Among persons with lung cancer, 29% of never smokers and 18% of ever smokers responded to the 2005–2010 interview, compared with 49% of never-smoking and 46% of ever-smoking respondents who did not have lung cancer, resulting in a lower cumulative incidence of lung cancer in the subcohort than the full cohort. Similar proportions of persons with and without bladder cancer responded to the 2005–2010 interview, leading to similar subcohort and full-cohort cumulative incidence estimates. A greater proportion of persons with prostate cancer responded than persons without. This led to an overestimate of the incidence of prostate cancer in the subcohort compared with the full cohort.

Standardized estimates of the risk ratio and risk difference from the subcohort and full cohort are shown in Table 5. Because of differential response proportions of lung cancer cases by smoking status, the subcohort risk ratio and risk difference were lower than the full-cohort estimates. Nondifferential response proportions of bladder and prostate cancer cases by smoking status led to subcohort risk ratio estimates similar to full-cohort estimates. The subcohort risk difference for ever smoking and bladder cancer was also similar to the full-cohort risk difference. For ever smoking and prostate cancer, the subcohort risk difference was similar to, but on the opposite side of the null and less precise than the full-cohort risk difference.

Illustrating bias reduction through IPSW

Differential response proportions for lung cancer cases by smoking status indicated the need for IPSW|E,**Z**,D to fully correct estimates. Application of IPSW|E,**Z**,D to the subcohort produced estimates of the risk ratio and risk difference that were similar to, but less precise than, full-cohort estimates (Table 5). Under these conditions, simpler IPSW|E,**Z** models were unable to fully correct subcohort estimates. The risk ratios and risk differences in the subcohort for bladder cancer and prostate cancer among ever smokers were already similar to full-cohort estimates, and therefore application of IPSW did not substantially alter results. However, application of IPSW|E,**Z** shifted the smoking–prostate cancer risk difference to the same side of the null as the full-cohort estimate.

DISCUSSION

Of farmers enrolled in the AHS (1993–1997), 46% responded to the 2005–2010 interview. The enrollment variables of age, state, education, race/ethnicity, marital status, smoking, and alcohol consumption were strongly associated with response. Personal use of pesticides and raising animals—variables often considered as exposures in AHS analyses—were not. We compared the incidence of 3 types of cancers in the full cohort with that in the subcohort of farmers responding to the 2005–2010 interview to identify conditions under which bias in measures of frequency and association would occur. Smoking was weakly associated with response. Only when the cancer outcome examined was strongly associated with response did we observe evidence of bias. When the outcome was not strongly associated with response, no evidence of bias was observed. These results suggest that where exposure and disease are not strongly associated with response, other AHS analyses based on the 2005–2010 interview population should provide good approximations of the associations that would have been obtained from the full AHS cohort.

Refusal and exclusion were the main reasons for nonresponse to the 2005–2010 AHS interview. Overall, 7% of enrolled farmers died before the interview. Because mortality accounted for a small proportion of nonrespondents, and with

Downloaded from https://academic.oup.com/aje/article/186/4/395/3806857 by Centers for Disease Control and Prevention user on 19 December 2023

Table 2. Associations Between Demographic and Lifestyle Variables Reported at Enrollment (1993–1997) and Response to the 2005–2010 Interview Among 52,394 Farmers, Agricultural Health Study, Iowa and North Carolina

Enrollment Characteristic	Original Enrollment (*n* = 52,394)	2005–2010 Response (*n* = 24,171)[a]		Nonresponse[a]				β[b]	SE	χ^2
				Refusal or Exclusion (*n* = 24,682)		Death (*n* = 3,541)				
	No. of Participants	No. of Participants	%	No. of Participants	%	No. of Participants	%			
Age, years										
<30	4,493	1,443	32	3,019	67	31	1	−0.63	0.04	235.12
30–39	12,141	5,205	43	6,788	56	148	1	−0.24	0.03	80.88
40–49	14,108	7,012	50	6,749	48	347	2	Referent		
50–59	11,155	5,913	53	4,491	40	751	7	0.22	0.03	61.96
60–69	7,768	3,781	49	2,723	35	1,264	16	0.11	0.03	11.75
70–79	2,487	784	32	836	34	867	35	−0.57	0.05	119.07
≥80	242	33	14	76	31	133	55	−1.61	0.22	51.75
State										
Iowa	31,876	15,760	49	14,572	46	1,544	5	Referent		
North Carolina	20,518	8,411	41	10,110	49	1,997	10	−0.23	0.02	103.24
Sex										
Female	1,362	674	49	616	45	72	5	0.15	0.08	5.90
Male	51,031	23,496	46	24,066	47	3,469	7	Referent		
Missing	1									
Race/ethnicity										
White	49,345	23,202	47	22,891	46	3,252	7	Referent		
Black	1,172	336	29	705	60	131	11	−0.38	0.08	23.79
Hispanic	523	206	39	275	53	42	8	−0.39	0.16	6.23
Other	288	93	32	163	57	32	11	−0.23	0.10	5.65
Missing	1,066									
Education										
Less than high-school diploma	5,224	1,840	35	2,580	49	804	15	−0.29	0.04	64.16
High school or equivalent	24,061	10,739	45	11,720	49	1,602	7	Referent		
Some college	12,119	5,988	49	5,608	46	523	4	0.23	0.02	94.89
College graduate or more	8,589	4,740	55	3,490	41	359	4	0.46	0.03	291.12
Missing	2,401									
Marital status										
Married/living as married	43,692	21,114	48	19,468	45	3,110	7	Referent		
Divorced/separated	2,299	756	33	1,437	63	106	5	−0.55	0.05	122.67
Widowed/never married	6,143	2,228	36	3,617	59	298	5	−0.24	0.03	56.41
Missing	260									
Smoking status, no. of pack years										
Never	26,690	13,031	49	12,476	47	1,183	4	Referent		
Former, <5	5,635	2,907	52	2,448	43	280	5	0.07	0.03	5.07
Former, 5–29	6,168	2,974	48	2,654	43	540	9	−0.06	0.03	3.78
Former, ≥30	2,831	1,275	45	1,060	37	496	18	−0.14	0.04	10.17
Current, <15	2,980	1,027	34	1,825	61	128	4	−0.35	0.04	63.14
Current, 15–44	3,618	1,405	39	1,912	53	301	8	−0.33	0.04	71.69
Current, ≥45	1,263	503	40	550	44	210	17	−0.37	0.06	33.92
Missing	3,209									

Table continues

Downloaded from https://academic.oup.com/aje/article/186/4/395/3806857 by Centers for Disease Control and Prevention user on 19 December 2023

Table 2. Continued

Enrollment Characteristic	Original Enrollment (*n* = 52,394)	2005–2010 Response (*n* = 24,171)[a]		Nonresponse[a]				β[b]	SE	χ^2
				Refusal or Exclusion (*n* = 24,682)		Death (*n* = 3,541)				
	No. of Participants	No. of Participants	%	No. of Participants	%	No. of Participants	%			
Alcohol consumption[c]										
None	16,837	7,936	47	7,255	43	1,646	10	Referent		
Light drinker	30,521	14,481	47	14,603	48	1,437	5	−0.10	0.02	19.30
Heavy drinker	1,113	375	34	708	64	30	3	−0.38	0.07	30.90
Missing	3,923	3,923								

Abbreviation: SE, standard error.

[a] The percentages shown are the proportion of respondents or nonrespondents (by death and refusals/exclusions) by the specified level of each enrollment characteristic.

[b] β coefficient is the change in log odds of response, comparing the specified level of each characteristic with the referent; estimated from the 47,007 participants with complete data on all variables.

[c] Heavy drinkers were defined as consuming ≥5 drinks on the same occasion on each of ≥5 days in the past 30 days. Light drinkers were those who reported consuming alcohol (≥1 drink on ≥1 day) during the past 12 months and who did not qualify as heavy drinkers (50).

a few exceptions a similar distribution of enrollment characteristics were observed for deaths and other nonrespondents, both groups were considered together as nonrespondents. Including deaths as contributing to nonresponse recognizes that those who died had an unobserved outcome status prior to death, and it reflects the main goal of these analyses: to determine whether we could obtain subcohort estimates similar to full-cohort estimates under conditions relevant to outcomes

Table 3. Associations Between Smoking Status Reported at Enrollment (1993–1997), Incident Cancer[a], and Selection into the 2005–2010 Interview Among Farmers, Agricultural Health Study, Iowa and North Carolina

Exposure	Original Enrollment (*n* = 47,007)	2005–2010 Response (*n* = 22,214)		Nonresponse				β[b]	SE	Wald χ^2
				Refusal or Exclusion (*n* = 21,840)		Death (*n* = 2,953)				
	No. of Participants[c]	No. of Participants[c]	%[d]	No. of Participants[c]	%[d]	No. of Participants[c]	%[d]			
Ever smoker										
No	25,169	12,389	49	11,702	46	1,078	4	Referent		
Yes	21,838	9,825	45	10,138	46	1,875	9	−0.16	0.02	65.21
Incident disease										
Lung cancer										
No	46,476	22,106	48	21,692	47	2,678	6	Referent		
Yes	507	100	20	141	28	266	52	−1.19	0.11	108.12
Bladder cancer										
No	46,706	22,068	47	21,750	47	2,888	6	Referent		
Yes	249	127	51	73	29	49	20	0.20	0.13	2.32
Prostate cancer										
No	43,450	20,324	47	20,540	47	2,586	6	Referent		
Yes	2,044	1,162	57	682	33	200	10	0.30	0.05	39.97

Abbreviation: SE, standard error.

[a] Incidence data (for lung, bladder, and prostate cancer) were obtained from state cancer registries.

[b] β coefficient is the change in log odds of response, comparing the specified level of each characteristic with the referent adjusted for age, state, race/ethnicity, education, marital status, and alcohol consumption.

[c] Numbers do not sum to totals because of the exclusion of prevalent cancer cases (lung cancer: *n* = 24; bladder cancer: *n* = 52; prostate cancer: *n* = 361).

[d] Proportion of respondents or nonrespondents (by death and refusals/exclusions) was relative to the specified level of each enrollment characteristic.

Downloaded from https://academic.oup.com/aje/article/186/4/395/3806857 by Centers for Disease Control and Prevention user on 19 December 2023

Table 4. Joint Distribution of Ever Smoking and Incident Cancer Among the Full Cohort ($n = 52,394$) and the Subcohort of Farmers Responding to the 2005–2010 Interview ($n = 24,171$), Agricultural Health Study, Iowa and North Carolina

Exposure	Case Response		Noncase Response		Cumulative Incidence per 1,000 Persons[b]	95% CI
	No. of Participants	%[a]	No. of Participants	%[a]		
Lung cancer						
Full cohort						
Never smoker	58	100	25,109	100	2.8	2.2, 3.7
Ever smoker	449	100	21,367	100	17.0	15.5, 18.7
Subcohort						
Never smoker	17	29	12,370	49	1.5	0.9, 2.5
Ever smoker	83	18	9,736	46	6.9	5.5, 8.6
Bladder cancer						
Full cohort						
Never smoker	67	100	25,087	100	3.1	2.4, 4.0
Ever smoker	182	100	21,619	100	7.4	6.3, 8.6
Subcohort						
Never smoker	37	55	12,347	49	3.3	2.4, 4.7
Ever smoker	90	50	9,721	45	8.3	6.6, 10.5
Prostate cancer						
Full cohort						
Never smoker	995	100	23,258	100	47.0	4.41, 50.0
Ever smoker	1,049	100	20,192	100	45.4	42.7, 48.3
Subcohort						
Never smoker	578	58	11,341	49	55.6	51.2, 60.3
Ever smoker	584	56	8,983	44	56.7	52.2, 61.7

Abbreviation: CI, confidence interval.

[a] Proportion of farmers responding in the respective exposure-disease category.

[b] Cumulative incidence was estimated from log-binomial and linear-binominal models with inverse probability of exposure weights applied. It is presented per 1,000 persons.

reported by 2005–2010 interview respondents. Estimating separate weights for total or cause-specific deaths, excluding individuals who experienced a competing event (e.g., death) prior to follow-up, or employing alternate methods to handle such issues as competing events may be more appropriate in other situations (30, 31).

Associations between enrollment factors and response were similar to those observed elsewhere. Researchers have previously observed nonresponse at follow-up activities associated with younger and older age (32–36), male sex (34, 35, 37), race/ethnicity other than non-Hispanic white (35), lower levels of education (7, 32, 35, 36), and marital status other than married (32, 36–38). Smoking (32, 35, 36, 38, 39) and heavy alcohol consumption (36, 39) have also been consistently associated with nonresponse in previous work. Abstention from alcohol has also been previously associated with attrition (36), but this was not associated with nonresponse in our analysis. Finally, other researchers have observed that persons reporting general poor health (33) or chronic illness (39) participate less in study activities. Although not identified as important to include in weight-generation models, we observed lower response proportions for farmers reporting diagnosed heart disease, diabetes, or Parkinson disease. Lower response proportions were not observed for farmers reporting other chronic conditions (e.g. asthma, other chronic lung disease, depression) possibly because these conditions rarely lead to rapid mortality or severe disability, as some of the other noted conditions can.

Relationships between enrollment characteristics and response were also consistent with factors associated with participation in previous AHS activities, including the 1999–2003 interview, with a few notable exceptions (18, 40, 41). In the 1999–2003 interview, nonresponse decreased with increasing age; however, that investigation excluded those who died prior to follow-up. This exclusion, coupled with continued aging of the cohort, may partially account for the difference. Response to the 1999–2003 interview was greater for individuals reporting illness at enrollment (18). We did not observe this, possibly because farmers reporting a diagnosis at enrollment have had more time to experience complications that could interfere with being interviewed. However, a greater proportion of incident prostate cancer cases than noncases responded to the 2005–2010 interview. This could indicate that disease-related interest in continued study participation may be operating along with disease-related reasons for nonresponse.

Although specific farm activities were not strongly associated with response, similar to the previous investigation (18),

Downloaded from https://academic.oup.com/aje/article/186/4/395/3806857 by Centers for Disease Control and Prevention user on 19 December 2023

Table 5. Standardized Risk Ratio and Risk Difference Estimates[a] Quantifying the Association Between Ever Smoking and Specified Incident Cancer Outcomes Among the Full Cohort and the Subcohort of Farmers Participating in the 2005–2010 Follow-up Interview, Agricultural Health Study, Iowa and North Carolina

Cancer Type	Full Cohort		Subcohort		Subcohort With IPSW|E,Z,D[b]		Subcohort With IPSW|E,Z[c]	
	RR	95% CI	RR	95% CI	RR	95% CI	RR	95% CI
Lung cancer	5.99	4.50, 8.00	4.60	2.68, 7.92	6.73	3.77, 12.00	4.84	2.75, 8.50
Bladder cancer	2.38	1.75, 3.24	2.51	1.66, 3.81	2.53	1.67, 3.85	2.55	1.68, 3.88
Prostate cancer	0.97	0.88, 1.06	1.02	0.91, 1.15	0.93	0.82, 1.05	0.98	0.87, 1.11
	RD	**95% CI**	**RD**	**95% CI**	**RD**	**95% CI**	**RD**	**95% CI**
Lung cancer	14.2	12.40, 16.00	5.40	3.70, 7.10	13.80	9.90, 17.70	5.40	3.70, 7.10
Bladder cancer	4.3	2.90, 5.70	5.00	2.80, 7.30	4.40	2.50, 6.40	4.80	2.70, 6.90
Prostate cancer	−1.6	−5.70, 2.50	1.20	−5.40, 7.80	−3.60	−9.30, 2.21	−1.00	−7.50, 5.50

Abbreviations: CI, confidence interval; IPSW, inverse probability of selection weight; RD, risk difference; RR, risk ratio.

[a] Risk ratios were estimated from log-binomial models, and risk differences were estimated from linear-binomial models.

[b] Risk ratios and risk differences with the application of inverse probability of selection weights conditional on exposure (E), covariates (**Z**), and disease (D).

[c] Risk ratios and risk differences with the application of inverse probability of selection weights conditional on exposure (E) and covariates (**Z**).

higher proportions of farmers reporting personal pesticide use and animal production at enrollment responded to the 2005–2010 interview compared with their counterparts. This may indicate that farmers actively engaged in these activities have a stronger interest in participating in a study of health outcomes associated with such exposures. Alternatively, farmers engaged in these activities may be healthier than those who are not (42, 43).

Researchers have previously used simulation studies and directed acyclic graph theory to investigate the potential influence of attrition in cohort studies (4, 6, 12, 44, 45). However, only a few previous studies have done so in a real-world setting (13, 46). Here, we have illustrated that associations estimated from those who remain under study may be biased under certain conditions—namely, when exposure and outcome were associated with each other and with response. Under these conditions the subcohort risk difference and risk ratio underestimated the full-cohort values. Although other conditions may also produce biased results, when the exposure and outcome were not strongly associated with response, subcohort risk difference and risk ratio estimates were similar to estimates from the full cohort.

When both the exposure and outcome were associated with response, IPSW|E,**Z**,D moved subcohort estimates toward full-cohort estimates; simpler weights estimated using only exposure and covariates did not. These observations were more evident when considering the risk difference for the smoking–lung cancer association compared with the risk ratio. The smoking–lung cancer risk ratio estimated using IPSW|E,**Z**,D was closest to, but overestimated, the full-cohort risk ratio. The overestimate may be a result of imprecision, or it is possible that our models did not fully account for selection.

When the outcome was not associated with response, IPSWs were unnecessary. Collectively, these observations align with findings from simulation studies and theoretical examples (4, 11, 12, 44). Further, these findings support previous conclusions that understanding associations present in the cohort is critical to characterizing the potential for bias and selecting appropriate statistical methods to mitigate resulting bias (46).

The findings reported here are specifically relevant for studies of agricultural exposures and outcomes reported only by 2005–2010 AHS interview respondents, but they may have broader application. First, applying pesticides and raising animals at enrollment were not strongly associated with response to the 2005–2010 interview. Second, many of the outcomes self-reported by 2005–2010 interview respondents (e.g., allergy, chronic obstructive pulmonary disease, asthma, arthritis) do not have high rates of rapid mortality or disability soon after diagnosis and therefore should not be strongly associated with response. If these assumptions are correct, selection bias should not strongly influence estimates of the association between farming exposures and many of the self-reported outcomes when analyses are limited to the 2005–2010 interview respondents.

Although these results and previous AHS analyses (47) indicate that IPSW|E,**Z** may not be necessary or effective at reducing bias, application of IPSW|E,**Z** may be useful when the exposure of interest is more strongly associated with response, as demonstrated using simulations and other cohorts, some with similar attrition to the AHS (35, 38, 48, 49). Further, application of IPSWs may be useful to confirm no substantial change in results when applied. To that end, the factors associated with response reported here can inform construction of weights for future AHS analyses restricted to 2005–2010 interview respondents.

Although we were able to evaluate associations between many enrollment characteristics and response, it is possible that other unmeasured characteristics were associated with response. Specifically, we were not able to examine associations between other incident medical conditions and response because this information was unavailable for nonrespondents. We were also unable to evaluate associations between updated agricultural work and response for the full cohort. Engagement

Downloaded from https://academic.oup.com/aje/article/186/4/395/3806857 by Centers for Disease Control and Prevention user on 19 December 2023

in farming activities after enrollment may be strongly associated with response to later interviews. In addition, it is possible that different criteria for selecting factors associated with attrition would identify an alternative set of variables to describe response. However, we were able to estimate weights that produced subcohort effect estimates similar to full-cohort estimates. Further, weights were robust to the inclusion of several additional variables in weight-generation models but sensitive to the exclusion of identified factors. These observations provide evidence that weight models were well-specified, one of the assumptions implicit in the use of inverse probability weighting (25). Finally, although this analysis provides relevant information about the potential for selection bias to influence future analyses of self-reported outcomes from the AHS 2005–2010 interview, the examples presented here represent specific sets of assumptions, and conditions relevant to future analyses should be evaluated on an individual basis.

Attrition is a common occurrence in many longitudinal studies. Beyond the AHS, this investigation might serve as an applied example supporting the findings of previous theoretical work regarding the influence of attrition on effect estimation from long-term, cohort studies that rely on self-reported information from participants over time. The findings reported here highlight that even in the presence of substantial attrition, it is important to consider the underlying associations present between relevant characteristics and attrition to understand the potential for selection bias and determine the utility of statistical methods to handle such bias.

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, North Carolina (Jessica L. Rinsky, David B. Richardson, Steve Wing, John D. Beard); Occupational and Environmental Epidemiology Branch, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Rockville, Maryland (Michael Alavanja, Laura E. Beane Freeman); Epidemiology Branch, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina (John D. Beard, Honglei Chen, Freya Kamel, Dale P. Sandler); National Institute for Occupational Safety and Health, Centers for Disease Control and Prevention, Morgantown, West Virginia (Paul K. Henneberger); and Department of Biological Sciences, Center for Human Health and the Environment, North Carolina State University, Raleigh, North Carolina (Jane A. Hoppin).

This work was supported by the Intramural Research Program of the National Institutes of Health, National Institute of Environmental Health Sciences (grant Z01-ES049030) and National Cancer Institute (grant Z01-CP010119). J.L.R. was supported by the National Institute of Occupational Safety and Health (award T42OH00867302).

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). Mention of any company or product does not constitute endorsement by National Institute for Occupational Safety and Health.

The authors thank the participants of the Agricultural Health Study for continued support and Stuart Long for assistance with data analysis. The authors also thank Dr. Leena Nylander-French, Dr. Kathleen Kreiss, and Dr. Alan Kinlaw for critical review of this work.

Conflict of interest: none declared.

REFERENCES

1. Nohr EA, Frydenberg M, Henriksen TB, et al. Does low participation in cohort studies induce bias? *Epidemiology*. 2006;17(4):413–418.
2. Doll R. Cohort studies: history of the method. I. Prospective cohort studies. *Soz Praventivmed*. 2001;46(2):75–86.
3. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008.
4. Kristman V, Manno M, Cote P. Loss to follow-up in cohort studies: how much is too much? *Eur J Epidemiol*. 2004;19(8): 751–760.
5. Kleinbaum DG, Morgenstern H, Kupper LL. Selection bias in epidemiologic studies. *Am J Epidemiol*. 1981;113(4):452–463.
6. Greenland S. Response and follow-up bias in cohort studies. *Am J Epidemiol*. 1977;106(3):184–187.
7. Streib GF. Participants and drop-outs in a longitudinal study. *J Gerontol*. 1966;21(2):200–209.
8. Schwartz S, Campbell UB, Gatto NM, et al. Toward a clarification of the taxonomy of “bias” in epidemiology textbooks. *Epidemiology*. 2015;26(2):216–222.
9. Hernán MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004;15(5): 615–625.
10. Soullier N, de La Rochebrochard E, Bouyer J. Multiple imputation for estimation of an occurrence rate in cohorts with attrition and discrete follow-up time points: a simulation study. *BMC Med Res Methodol*. 2010;10:79.
11. Daniel RM, Kenward MG, Cousens SN, et al. Using causal diagrams to guide analysis in missing data problems. *Stat Methods Med Res*. 2012;21(3):243–256.
12. Westreich D. Berkson’s bias, selection bias, and missing data. *Epidemiology*. 2012;23(1):159–164.
13. Twisk J, de Vente W. Attrition in longitudinal studies. How to deal with missing data. *J Clin Epidemiol*. 2002;55(4):329–337.
14. Van Beijsterveldt CE, van Boxtel MP, Bosma H, et al. Predictors of attrition in a longitudinal cognitive aging study: the Maastricht Aging Study (MAAS). *J Clin Epidemiol*. 2002; 55(3):216–223.
15. Wolke D, Waylen A, Samara M, et al. Selective drop-out in longitudinal studies and non-biased prediction of behaviour disorders. *Br J Psychiatry*. 2009;195(3):249–256.
16. Alavanja MC, Sandler DP, McMaster SB, et al. The Agricultural Health Study. *Environ Health Perspect*. 1996; 104(4):362–369.
17. Agricultural Health Study. About the study. https://aghealth.nih.gov/about/. Accessed August 8, 2013.
18. Montgomery MP, Kamel F, Hoppin JA, et al. 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(5):486–496.

Downloaded from https://academic.oup.com/aje/article/186/4/395/3806857 by Centers for Disease Control and Prevention user on 19 December 2023

19. Hoppin JA, Umbach DM, Long S, et al. Respiratory disease in US farmers. *Occup Environ Med*. 2014;71(7):484–491.
20. Agricultural Health Study. Questionnaires and study data. http://www.aghealth.nih.gov/collaboration/questionnaires.html. Accessed August 8, 2013.
21. Doll R, Hill AB. Smoking and carcinoma of the lung; preliminary report. *Br Med J*. 1950;2(4682):739–748.
22. Brennan P, Bogillot O, Cordier S, et al. Cigarette smoking and bladder cancer in men: a pooled analysis of 11 case-control studies. *Int J Cancer*. 2000;86(2):289–294.
23. Hickey K, Do KA, Green A. Smoking and prostate cancer. *Epidemiol Rev*. 2001;23(1):115–125.
24. Giovannucci E, Rimm EB, Ascherio A, et al. Smoking and risk of total and fatal prostate cancer in United States health professionals. *Cancer Epidemiol Biomarkers Prev*. 1999; 8(4 Pt 1):277–282.
25. Cole SR, Hernan MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol*. 2008; 168(6):656–664.
26. Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*. 2000;11(5): 561–570.
27. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11(5):550–560.
28. Brookhart MA, Schneeweiss S, Rothman KJ, et al. Variable selection for propensity score models. *Am J Epidemiol*. 2006; 163(12):1149–1156.
29. Beard JD, Hoppin JA, Richards M, et al. Pesticide exposure and self-reported incident depression among wives in the Agricultural Health Study. *Environ Res*. 2013;126:31–42.
30. Lau B, Cole SR, Gange SJ. Competing risk regression models for epidemiologic data. *Am J Epidemiol*. 2009;170(2):244–256.
31. Tchetgen Tchetgen EJ. Identification and estimation of survivor average causal effects. *Stat Med*. 2014;33(21):3601–3628.
32. Benfante R, Reed D, MacLean C, et al. Response bias in the Honolulu Heart Program. *Am J Epidemiol*. 1989;130(6): 1088–1100.
33. Boshuizen HC, Viet AL, Picavet HS, et al. Non-response in a survey of cardiovascular risk factors in the Dutch population: determinants and resulting biases. *Public Health*. 2006;120(4): 297–308.
34. Langley JD, Lilley R, Wilson S, et al. Factors associated with non-participation in one or two follow-up phases in a cohort study of injured adults. *Inj Prev*. 2013;19(6):428–433.
35. Littman AJ, Boyko EJ, Jacobson IG, et al. Assessing nonresponse bias at follow-up in a large prospective cohort of relatively young and mobile military service members. *BMC Med Res Methodol*. 2010;10:99.
36. Thomas MC, Walker M, Lennon LT, et al. Non-attendance at re-examination 20 years after screening in the British Regional Heart Study. *J Public Health Med*. 2002;24(4): 285–291.
37. Langhammer A, Krokstad S, Romundstad P, et al. The HUNT study: participation is associated with survival and depends on socioeconomic status, diseases and symptoms. *BMC Med Res Methodol*. 2012;12:143.
38. Alonso A, Segui-Gomez M, de Irala J, et al. Predictors of follow-up and assessment of selection bias from dropouts using inverse probability weighting in a cohort of university graduates. *Eur J Epidemiol*. 2006;21(5):351–358.
39. Goldberg M, Chastang JF, Zins M, et al. Health problems were the strongest predictors of attrition during follow-up of the GAZEL cohort. *J Clin Epidemiol*. 2006;59(11):1213–1221.
40. Tarone RE, Alavanja MC, Zahm SH, 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(2):233–242.
41. Engel LS, Rothman N, Knott C, et al. Factors associated with refusal to provide a buccal cell sample in the Agricultural Health Study. *Cancer Epidemiol Biomarkers Prev*. 2002; 11(5):493–496.
42. Fox AJ, Collier PF. Low mortality rates in industrial cohort studies due to selection for work and survival in the industry. *Br J Prev Soc Med*. 1976;30(4):225–230.
43. Arrighi HM, Hertz-Picciotto I. The evolving concept of the healthy worker survivor effect. *Epidemiology*. 1994;5(2): 189–196.
44. Kristman VL, Manno M, Cote P. Methods to account for attrition in longitudinal data: do they work? A simulation study. *Eur J Epidemiol*. 2005;20(8):657–662.
45. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*. 2003;14(3):300–306.
46. Weisskopf MG, Sparrow D, Hu H, et al. Biased exposure-health effect estimates from selection in cohort studies: are environmental studies at particular risk? *Environ Health Perspect*. 2015;123(11):1113–1122.
47. Beard JD, Umbach DM, Hoppin JA, et al. Pesticide exposure and depression among male private pesticide applicators in the Agricultural Health Study. *Environ Health Perspect*. 2014; 122(9):984–991.
48. Weuve J, Tchetgen Tchetgen EJ, Glymour MM, et al. Accounting for bias due to selective attrition: the example of smoking and cognitive decline. *Epidemiology*. 2012;23(1):119–128.
49. Gottesman RF, Rawlings AM, Sharrett AR, et al. Impact of differential attrition on the association of education with cognitive change over 20 years of follow-up: the ARIC neurocognitive study. *Am J Epidemiol*. 2014;179(8):956–966.
50. National Institute on Alcohol Abuse and Alcoholism. Drinking levels defined. http://www.niaaa.nih.gov/alcohol-health/overview-alcohol-consumption/moderate-binge-drinking. Accessed February 18, 2015.

Downloaded from https://academic.oup.com/aje/article/186/4/395/3806857 by Centers for Disease Control and Prevention user on 19 December 2023