Effect of Proxy-reported Smoking Status on Population Estimates of Smoking Prevalence

Andrew Hyland,1 K. Michael Cummings,1 William R. Lynn,2 Donald Corle,3 and Carol A. Giffen4

The use of proxy respondents in surveys designed to provide population estimates of smoking prevalence offers an inexpensive way to obtain these data. The accuracy of this information is examined in analyzing data from tobacco use surveys of adults conducted in 22 North American communities as part of the National Cancer Institute's Community Intervention Trial for Smoking Cessation. Proxy-reported smoking status was obtained in a cross-sectional telephone survey conducted from August 1993 to January 1994 (n = 99,682). Self-reported smoking status was obtained from an in-depth interview of a sample of the respondents aged 25-64 years enumerated from the telephone survey (n = 31,417). Discrepancy rates were calculated by comparing the proxy-reported and self-reported smoking statuses of a given individual (n = 10,226). In both surveys, respondents were categorized as current smokers (those who currently smoke and have smoked at least 100 cigarettes in their lifetime), recent quitters (≤8 years since cessation), long-term quitters (>8 years since cessation), and never smokers. The overall discrepancy rate between the self-report and the proxy report was 5.4%. Self-respondents who were black, Hispanic, Asian, recent quitters, or aged 25-34 years were more likely to have inconsistent proxy reports. The authors estimate that the screener interview underestimated the true smoking prevalence by 0.1% when they corrected for smoking status discrepancies. These results confirm that proxy-reported smoking status is an accurate and effective means to monitor populationwide smoking prevalence of adults. Am J Epidemiol 1997;145:746-51.

Smoking is the single largest cause of preventable death in the United States today. To monitor trends in smoking behavior, it is important to implement surveillance systems to track smoking prevalence (1, 2). As with any surveillance system, it is necessary to consider the optimal strategy with regard to cost and validity.

Existing surveys, such as the Current Population Survey conducted by the Bureau of the Census, obtain information on all household members from a single telephone respondent, i.e., use proxy respondents. Surveys that have attempted to interview each adult household member have had poor response rates (3). Therefore, surveys that use a proxy to elicit information about other household members give a more representative sample as well as increase the effective sample size for only a marginal increase in the cost per interview.

Populationwide smoking prevalence estimates were obtained by using proxy respondents from the National Cancer Institute's Community Intervention Trial for Smoking Cessation (COMMIT) Final Prevalence Survey conducted in 1993. A sample of adults who had their smoking status enumerated from this screener interview by a proxy respondent were selected to participate in an in-depth interview. Therefore, we have the opportunity to examine the relation between the proxy-reported smoking status of individuals from the screener interview and their self-reported smoking status from the in-depth interview. This study uses these data to answer the following three questions: 1) How accurate is proxy information on smoking status? 2) What factors are related to inconsistent proxy-reported smoking status? and 3) What is the impact of this misclassification on the overall estimate of smoking prevalence in a population?
MATERIALS AND METHODS

The COMMIT Study

The COMMIT Study was a randomized controlled trial conducted at the community level to test the effectiveness of a multifaceted intervention to help adult smokers achieve and maintain cessation (4). The study involved 11 matched pairs of communities: 10 pairs of communities in the United States and one pair in Canada. The design and primary outcomes of the COMMIT Study have been described elsewhere (5–7).

The data presented in this paper come from a cross-sectional telephone survey conducted in 1993 at each of the 22 COMMIT sites. These communities included Bellingham and Longview/Kelso, Washington; Albany/Corvallis and Medford/Ashland, Oregon; Hayward and Vallejo, California; Santa Fe and Las Cruces, New Mexico; Cedar Rapids and Davenport, Iowa; Raleigh and Greensboro, North Carolina; Paterson and Trenton, New Jersey; Lowell and Fitchburg/Leominster, Massachusetts; Yonkers, New Rochelle, Utica, and Binghamton/Johnson City, New York; and Brantford and Peterborough, Ontario, Canada.

Data collection

A computer-assisted telephone interviewing system was used to administer the final prevalence survey in households that were screened by geographic boundaries and selected by a modified random digit dialing method. This approach was used to obtain representative samples of an average of 2,300 households within each of the 22 communities. The sample of telephone numbers was obtained from all numbers (both listed and unlisted) whose exchange prefixes covered the COMMIT study areas. The A. C. Nielsen Total Telephone Frame was used for telephone numbers in the United States; telephone numbers in Canada were provided by the Institute of Survey Research at York University, Ontario, Canada. COMMIT used a disproportionate stratified random digit dial sampling design with two strata per community. This type of design uses stratification based on how likely a telephone number is to be attached to a household, and it allocates the sample disproportionately among the strata to decrease cost. Furthermore, each household in the screener enumeration was assigned a sampling weight to account for both the sampling design and nonresponse and to allow for communitywide population estimates. These weights reflect the probability of a given household being included in the survey. Details on how the data were weighted are described elsewhere (7). All analyses presented in this paper use weighted data.

In each selected household (n = 51,832), an adult (age 18 years or older) was administered a screener questionnaire that obtained basic demographic information, including smoking status, for each household member age 18 years or older. Of the 99,682 subjects enumerated from the screener interview, 31,417 were selected to participate in the extended interview (50.5 percent smokers, 24.7 percent recent quitters, 12.2 percent long-term quitters, and 12.5 percent never smokers). A sample of this size was chosen to obtain sufficient numbers of heavy and light-to-moderate smokers to address COMMIT’s primary study objective to test the impact of a communitywide intervention to increase the rate of smoking cessation among smokers in the population (4, 8). Of the 31,417 individuals selected to do an extended interview, 16,599 were interviewed directly. Therefore, there was no proxy interview completed for these individuals, and they were dropped from our analysis. From the remaining 14,818 proxy-self pairs, subjects were dropped from the analysis if they were missing information on age or gender (n = 161; 1.1 percent), proxy-reported smoking status (n = 375; 2.5 percent), or self-reported smoking status (n = 4,056; 27.4 percent). With these deletions, the final sample available for analysis included 10,226 proxy-self pairs. Nonresponders were overrepresented by females, blacks, and people with low incomes. A proxy-self pair condition was met when a person selected for the in-depth interview had a member of their household provide their smoking status by proxy on the screener interview.

Of the 10,226 proxy-self pairs considered in this analysis, the in-depth interview identified 50.4 percent self-reported smokers, 22.1 percent self-reported recent quitters (≤8 years since cessation), 13.1 percent self-reported long-term ex-smokers (>8 years since cessation), and 14.4 percent self-reported never smokers. There is a disproportionate percentage of smokers in the sample because the survey was designed to oversample smokers. In addition, note that 53.6 percent of the people listed on the screener interview provided their own smoking status.

Smoking status

In the screener interview, the smoking status of a person was determined by responses to the following questions asked of the proxy respondent for each household member age 18 years or older: “As far as you know, has [person] ever been a cigarette smoker, that is, smoked at least 100 cigarettes in his/her lifetime?” If the answer was no, the listed household member was classified as a never smoker. If the answer was yes, the respondent was asked, “Does [person] smoke cigarettes now?” Those whose
proxy respondents answered this question affirmatively were classified as current smokers, and those whose proxy respondents answered negatively were asked the question, "Has [person] quit within the last eight years as far as you know?" Those whose proxy respondents answered yes were classified as recent quitters, while those whose proxy respondents answered no were classified as long-term quitters.

Each person selected for the in-depth interview was asked the same set of questions to ascertain smoking status. Therefore, we have the ability to verify the smoking status of people who were selected for the in-depth interview and whose smoking status was reported by a proxy respondent from the same household.

**Analytic methods**

For each category of self-reported smoking status, the weighted proxy-reported percentages of current smokers, recent quitters, long-term quitters, and never smokers were calculated. In addition, the likelihood of a discrepancy was modeled using logistic regression. A discrepancy is defined as a proxy-reported smoker who self-reported being a current nonsmoker or a proxy-reported nonsmoker who self-reported being a current smoker.

To calculate the effect of discrepancies on the overall estimate of smoking prevalence, discrepancy proportions for each category of individuals classified according to smoking status from the screener interview were estimated from the survey data. The four specific categories of discrepancies are proxy-reported smokers who reported themselves to be nonsmokers ($D_1$), proxy-reported recent quitters who reported themselves to be smokers ($D_2$), proxy-reported long-term quitters who reported themselves to be smokers ($D_3$), and proxy-reported never smokers who reported themselves to be smokers ($D_4$) (9). Note that $D_1$, $D_2$, $D_3$, and $D_4$ are weighted discrepancy proportions.

The estimated smoking prevalence from the screener survey is given below:

$$ p = \frac{\sum_{i=1}^{N} w_i \times S_i}{\sum_{i=1}^{N} w_i} $$

where $w_i$ is the screener weight given to a particular person and $s_i = 1$ if the respondent is a current smoker and 0 otherwise. $N$ is the total number of adults from the screener interview.

The "discrepancy-corrected" smoking prevalence estimate from the screener interview is given as:

$$ p = \left( \sum_{i=1}^{n_0} w_i \times S_i \right) + \left( 1 - D_1 \right) \sum_{i=1}^{n_1} w_i + D_2 \sum_{i=1}^{n_2} w_i + D_3 \sum_{i=1}^{n_3} w_i + D_4 \sum_{i=1}^{n_4} w_i / \sum_{i=1}^{N} w_i $$

where $n_0 =$ number of respondents who reported their own smoking status on the screener, $n_1 =$ number of proxy-reported current smokers, $n_2 =$ number of proxy-reported recent quitters, $n_3 =$ number of proxy-reported long-term quitters, and $n_4 =$ number of proxy-reported never smokers (9).

**RESULTS**

Table 1 shows the bivariate relation of the proxy-reported smoking status from the screener interview to the self-reported smoking status from the in-depth interview. The percentages shown are weighted column percentages of the distribution of proxy classification for each category of self-classification. Self- and proxy reports of current smoking agreed 94.6 percent of the time. The largest inconsistencies occurred among long-term quitters, who were correctly classified by proxy as such only 44.2 percent of the

### TABLE 1. Agreement of proxy-reported smoking status and self-reported smoking status from the 1993 COMMIT* survey†

<table>
<thead>
<tr>
<th>Proxy-reported smoking status</th>
<th>Current smoker</th>
<th>Quitted ≤8 years previously</th>
<th>Quitted &gt;8 years previously</th>
<th>Never-smoker</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No.</td>
<td>%</td>
<td>No.</td>
<td>%</td>
<td>No.</td>
<td>%</td>
</tr>
<tr>
<td>Current smoker</td>
<td>4,949</td>
<td>95.8</td>
<td>214</td>
<td>10.4</td>
<td>29</td>
</tr>
<tr>
<td>Quitted ≤8 years previously</td>
<td></td>
<td>190</td>
<td>3.6</td>
<td>1,965</td>
<td>62.7</td>
</tr>
<tr>
<td>Quitted &gt;8 years previously</td>
<td></td>
<td>13</td>
<td>0.2</td>
<td>92</td>
<td>4.2</td>
</tr>
<tr>
<td>Never-smoker</td>
<td></td>
<td>21</td>
<td>0.4</td>
<td>60</td>
<td>2.6</td>
</tr>
<tr>
<td>Total</td>
<td>5,173</td>
<td>100.0</td>
<td>2,331</td>
<td>100.0</td>
<td>1,334</td>
</tr>
</tbody>
</table>

* COMMIT, Community Intervention Trial for Smoking Cessation.
† Percentages are weighted by household screener weights (see Materials and Methods).
‡ Overall agreement on current smoking status = 94.6.
time. We found 315 proxy-reported smokers who self-reported being a nonsmoker and 224 proxy-reported nonsmokers who self-reported smoking. When weighted, these account for 3.3 and 2.1 percent, respectively, nearly negating the error associated with each.

To examine the simultaneous effects of potential confounders on the likelihood of a proxy-self pair reporting discrepancies, a logistic regression model was fitted. The independent variable is whether the proxy correctly classified the smoking status of the household member. Dependent variables included in the model were the number of adults in the household as well as the age, gender, education, race, and smoking status of the self-respondent. Table 2 shows the distributions of various factors related to the household and self-respondents. Long-term quitters had the lowest discrepancy rates. The highest rates were found for those who were black, Hispanic, and Asian; those with annual household incomes above $10,000; persons aged 25–34 years; and recent quitters. These factors remained statistically significant even when we

| TABLE 2. Proxy-self discrepancies in subgroups defined by various factors related to the household and proxy of self-respondents from the 1993 COMMIT* surveys (n = 10,226) |
|---------------------------------|-----------------|------|--------|---|
| Characteristic                  | No. of pairs†   | %    | % misclassified | p value‡ |
| Overall                         | 10,226          | 100.0| 5.4            |   |
| Sex of self-respondent          |                 |      |                |   |
| Male                            | 5,978           | 58.5 | 5.7            | Reference |
| Female                          | 4,248           | 41.5 | 5.0            | 0.174 |
| Age of self-respondent (years)  |                 |      |                |   |
| 25–34                           | 3,102           | 30.3 | 6.8            | Reference |
| 35–44                           | 3,127           | 30.6 | 5.0            | 0.004 |
| 45–54                           | 2,528           | 24.7 | 4.7            | 0.003 |
| 55–64                           | 1,469           | 14.4 | 4.2            | 0.016 |
| Race/ethnicity of self-respondent|                |      |                |   |
| White                           | 5,777           | 65.7 | 4.1            | Reference |
| Black                           | 639             | 7.6  | 6.1            | <0.001 |
| Hispanic                        | 828             | 9.4  | 7.7            | <0.001 |
| Asian                           | 191             | 2.2  | 7.6            | 0.002 |
| American Indian                 | 123             | 1.4  | 0.0            | 0.631 |
| Canadian                        | 1,223           | 13.9 | 4.8            | 0.822 |
| Other                           | 18              | 0.2  | 0.0            | 0.885 |
| Education of self-respondent (years) |            |      |                |   |
| <12                             | 1,390           | 16.1 | 4.4            | Reference |
| 12                              | 1,708           | 19.8 | 5.2            | 0.321 |
| 13–15                           | 3,634           | 42.1 | 4.3            | 0.587 |
| ≥16                             | 1,893           | 21.9 | 5.0            | 0.025 |
| Gross household income          |                 |      |                |   |
| <$10,000                        | 510             | 6.3  | 3.0            | Reference |
| ≥$10,000                        | 7,565           | 93.7 | 4.7            | 0.040 |
| No. of adults in household      |                 |      |                |   |
| 2                               | 6,641           | 64.9 | 5.1            | Reference |
| 3                               | 2,319           | 22.7 | 5.1            | 0.923 |
| 4                               | 930             | 9.1  | 6.7            | 0.292 |
| ≥5                              | 336             | 3.3  | 10.5           | 0.233 |
| Smoking status of self-respondent|               |      |                |   |
| Current                         | 5,173           | 50.6 | 4.2            | Reference |
| Recent quitter (≤8 years)        | 2,331           | 22.8 | 10.4           | <0.001 |
| Long-term quitter (>8 years)     | 1,334           | 13.0 | 2.1            | <0.001 |
| Never-smoker                    | 1,388           | 13.6 | 5.1            | 0.136 |

* COMMIT, Community Intervention Trial for Smoking Cessation.
† For some variables, the number of proxy-self pairs does not total 10,226 due to missing data.
‡ p value from logistic regression model that adjusts for all other variables in the table.
performed logistic regression analysis to control for the other factors.

To estimate the effect of reporting errors by the proxy on the overall estimate of smoking prevalence, we estimated the four error rates, $D_1$, $D_2$, $D_3$, and $D_4$, to be 6.4, 6.5, 1.2, and 1.5 percent, respectively, from the in-depth interview. The corrected estimated smoking prevalence is 21.9 percent, while the uncorrected estimate is 21.8 percent. Therefore, we estimate that the use of proxy information in the COMMIT Study underestimated the true prevalence of smoking by 0.1 percent.

**DISCUSSION**

The use of proxy respondents in a population survey can dramatically increase statistical power within a fixed budget and can reduce costs of information collection. We found that the loss of information associated with using proxy respondents to report the smoking status of adult household members was minimal. The two types of misclassification of smoking status were infrequent and of similar magnitude. Therefore, the small effect of each tended to cancel out the other.

Some characteristics of self-respondents were significantly associated with higher rates of discrepancy. We found that race, age, higher income, and status as a recent quitter were all associated with elevated proxy misclassification levels. Concern for the accuracy of such proxy information could arise when the population under study possesses characteristics that are related to increased discrepancy rates. Smoking status misclassification can be of two types: 1) self-reported nonsmokers whose proxies report them as current smokers, and 2) self-reported smokers whose proxies report them as nonsmokers. However, we looked at factors associated with each direction of smoking status misclassification and found that many of the same factors were related in similar ways for both types of misclassification. Therefore, although the overall discrepancy rate may be higher in some populations, the net effect will still be predicted to be relatively small due to the cancellation effect noted above. For example, among those between ages 25 and 34 years, the rate of smoking status misclassification was 6.8 percent, but the prevalence rate from the screener survey underestimated the corrected prevalence rate by only 1.1 percent. Those aged 35–44 years had a smoking status misclassification rate of 5.0 percent; however, the prevalence rate from the screener interview underestimated the corrected prevalence by only 0.6 percent.

We assumed that the self-reported smoking status was the truly correct one. However, other studies have reported that nearly 30 percent of self-reported quitters who participated in smoking cessation programs had positive biochemical tests that indicated they had recently smoked (10–12). In this study, only 10.4 percent of recent quitters were classified by their household proxy as smokers. This is lower than the number cited in other studies (9); however, since our definition of recent quitter includes a much longer period of cessation, it may be easier for a proxy respondent to identify a household member as a recent quitter. In addition, there was a time lag between the screener interview and the in-depth interview for a number of respondents. Therefore, a person could be correctly classified at both points in time but still be counted as a discrepant case if he or she quit or initiated smoking in that time. The effect of this type of error on the overall smoking prevalence rate should be relatively small because the category of recent quitters and initiators comprises a minority of all adults.

To assess the smoking problem on a population level, it may also be important to obtain information such as age when smoking started, amount smoked, and other factors. There is some debate about whether this type of more detailed information can be obtained accurately from a proxy respondent (13–17). However, the evidence presented here as well as in other studies further confirms that proxy respondents give accurate information on smoking status for adult household members.

Therefore, the approach of having one household member report the smoking status for all adults in the household is a reliable, cost-efficient method of assessing the smoking prevalence in a population.

**ACKNOWLEDGMENTS**

Supported by the National Cancer Institute and the Robert Wood Johnson Foundation.

The authors are indebted to their colleagues at the 11 COMMIT research sites, the COMMIT coordination center, and Information Management Services, Inc., and to the staff at the National Cancer Institute who planned and coordinated the massive data collection activities of the COMMIT study.

**REFERENCES**


