Little is known about the comparability of postal and telephone survey modes in epidemiology. A cross-sectional, population-based study (n = 25,000) of lung disease was performed in 1998–1999 in two regions of Norway. Initial surveying was done by postal questionnaire. A 1% random sample (n = 171) of previous postal responders were resurveyed by telephone or cellular contact. The response rate was 67% on the telephone/cellular interview. Fewer incomplete answers were given by telephone than by mail. A lower prevalence was found by telephone for morning cough and exposure to passive smoking at work or home. Reproducibility was high for asthma, hay fever, wheezing, and attacks of breathlessness. Moderate reproducibility was seen for symptoms of chronic bronchitis. Reproducibility was low for indoor and work environment, although it was high for early life factors. Concordance coefficients were high for all continuous measures such as height, body weight, and pack-years. The authors conclude that the comparability between the postal and the telephone survey modes was good. The telephone survey mode gave more complete information. Survey mode may have a moderate effect on study results, depending on the specific questions asked. Am J Epidemiol 2002;155:572–6.

Common data collection instruments in epidemiology are postal questionnaires and telephone interviews. The comparability of these two survey modes needs to be studied. First of all, different researchers in any given epidemiologic field may be using different survey methods, so comparability of results is desirable (1).

Second, response rates are falling (2), so that surveys using both survey modes may become more common (3, 4). For example, an initial postal survey may be followed by a telephone survey of nonresponders to the postal survey. This will increase response rates, thus reducing nonresponse bias. However, there is a risk of introducing another bias, if responses to one survey method are systematically different from those of the other survey method.

There are several ways of comparing mail and telephone survey methods (1). Some have randomized half the population to the postal method and the other half to the telephone method (5–7). Others have compared postal responders with a telephone survey of postal nonresponders (8–10). The population that you reach with each survey method may be different, so these studies are unable to separate the effect of the survey method itself from the nonresponse bias due to survey method. Few epidemiologic studies have directly compared responses from the same subjects when surveyed by both postal questionnaire and telephone interview (11, 12).

The aim of the present study was to investigate the comparability of postal and telephone survey modes, through a telephone reinterview of responders to a postal survey, with respect to prevalence estimates, completeness, and reproducibility.

MATERIALS AND METHODS

The design was a cross-sectional study of the adult population in two regions of Norway, the capital Oslo and the county of Hordaland. The sampling frame was inhabitants listed in the Central Population Register who were aged 15–70 years on December 31, 1997. Random samples of 20,000 (Oslo) and 5,000 (Hordaland) inhabitants were selected.

Fieldwork

The study started with the mailing of a postal questionnaire. After 3 and 8 weeks a reminder letter was sent out. Four months after the beginning of the study, a telephone follow-up was initiated for a 1 percent random sample of
previous postal responders \((n = 171)\). Telephone numbers for the follow-up sample were located using a direct-mail firm (D. M. Huset, Oslo, Norway), matching on name, date of birth, and address. Six months after the beginning of the study, computer-assisted telephone interviews of these previous postal responders were performed by trained operators. A standardized interview protocol was followed. Nine months after the study began, cellular telephone numbers were located for those who did not have a land-line telephone. This was done through a search of the cellular phone companies’ customer databases. A similar cellular telephone interview was performed.

All mailings included an addressed envelope and return postage. The mailing and processing of the questionnaires were performed by Statistics Norway, the Norwegian government’s statistics and survey agency.

The sample size was chosen to provide reasonably precise estimates of reliability and to detect only large differences in prevalence. With 115 responders, a kappa of 0.70, and symmetric prevalences of 10 percent and 25 percent, the widths of the 95 percent confidence intervals would be 0.16 and 0.11, respectively. For a kappa of 0.90, they would be 0.10 and 0.07.

**Questionnaire**

The symptom questions were taken from a modified translation of the questionnaire by the British Medical Research Council’s committee on chronic bronchitis (13, 14), and they have been compared with a direct translation (15). A one-sheet questionnaire designed for self-completion was used. It contained 67 questions on respiratory symptoms, cardiorespiratory diagnoses, smoking habits, and risk factors. Thirty-five of these questions were meant to be answered by all participants. Three questions on heart disease and tuberculosis were not analyzed. Smokers were defined as daily smokers at the time of the study. Former smokers were persons who had smoked daily and had given it up. Pack-years were analyzed among smokers and former smokers only. The questionnaire used in the present study is available on the following Web site (http://www.brogger.no/ob98tlf/), as is an expanded version of the table, including all cell counts.

The study was approved by the regional ethics committee and the Norwegian Data Inspectorate.

**Statistical methods**

For comparisons between the survey modes, prevalences and odds ratios with 95 percent confidence intervals were estimated. Agreement and Cohen’s kappa (16, 17) were calculated for reproducibility. For ordinal data, Wilcoxon’s signed rank test was used to compare distributions, and Cohen’s unweighted kappa was computed for reproducibility. For continuous data, the confidence intervals for medians were computed and Lin’s concordance coefficient (18) was computed for reproducibility. For height and weight, a normal distribution was assumed. Confidence intervals for the mean number of missing answers on each questionnaire were computed using the negative binomial distribution (19). A global test of the effect of survey mode on outcomes was performed. The difference between survey modes for each outcome was summed for each person. The median of this sum across all subjects was tested with Wilcoxon’s signed rank test. The influence of age, gender, and region on the chance of postal response, having a telephone, and telephone response was assessed by logistic regression. Cases with missing answers with one or both survey modes were excluded from analyses involving that answer. All data were analyzed using Stata 7.0 software (20).

**RESULTS**

The postal response rate was 68 percent, leaving 17,083 responders available for selection for the follow-up, from which 171 were randomly selected. A land-line telephone was located for 141 subjects (83 percent), and a cellular telephone was located for 15 of the 30 remaining subjects, giving an overall coverage of 91 percent. A telephone interview was achieved with 115 of the available subjects (74 percent). Of the 56 subjects for whom no interview was achieved, 15 did not have a land-line telephone or a cellular telephone, seven were not reached or had moved, 13 were reached but refused participation, and 21 had other or unknown reasons for nonresponse.

**Completeness**

Of the 35 questions and one derived smoking habit variable, the mean number of incomplete questions was 2.8 (95 percent confidence interval (CI): 2.2, 3.6) with the postal survey and 0.6 (95 percent CI: 0.4, 0.8) with the telephone survey \((p \text{ for the difference } < 0.001)\). Few questions had any appreciable missing answers with the telephone survey mode.

In the postal survey, 17 questions had more than 5 percent missing answers. All symptoms had less than 5 percent missing answers except “episodes of phlegm or cough” (7 percent missing). The common diagnoses of asthma and bronchitis had 8 percent and 6 percent missing, while emphysema had 16 percent. Questions about eczema and hay fever had 28 and 31 percent missing, respectively. Other questions that had high rates of missingness were early life exposures (7–14 percent) and asthma in the family (14–23 percent).

**Prevalences and means**

The largest significant difference in prevalence between survey modes was for morning cough (table 1). Relatively large differences were observed for questions on dyspnea and for indoor environment questions, but with no consistent pattern in any direction.

Among ordinal variables, significantly less passive smoking at home was reported in the postal survey \((p = 0.001)\). Similarly, less passive smoking at work was reported in the postal survey, with any exposure at 55 percent versus 22 percent \((p < \ldots)\).
between survey modes (siblings with asthma, there was no significant difference (0.35 cm (95 percent CI: –0.03, 0.73; increase in height from postal to telephone survey mode was there were small changes of little significance. The mean agreement low were morning and day cough, dyspnea grade (>95 percent) agreement. Questions with both kappa and deformed floor, all had low kappas of <0.5 but high (>95 percent) agreement. Questions with both kappa and agreement low were morning and day cough, dyspnea grade 2, symptoms at work, and moisture at home. Other questions had intermediate kappas and agreements. For ordinal variables, the kappas were good for education (kappa, 0.80) and smoking (kappa, 0.84) and low for passive smoking (kappa, 0.40 at home; kappa, 0.33 at work) and episodes of cough or phlegm (kappa, 0.24).

For continuous variables, concordance coefficients were very high (>0.95 or better) for height, weight, and number of siblings; high for pack-years (0.82; 95 percent CI: 0.69, 0.94), and moderate for number of siblings with asthma (0.76, 95 percent CI: 0.68, 0.84).

**DISCUSSION**

We have shown that there were more missing answers using the postal survey mode than the telephone mode. Prevalences were similar for most questions, except for questions on cough, dyspnea, episodes of cough and phlegm, and passive smoking. Reproducibility was good for continuous variables, diagnoses, and asthma in the family but poor for dyspnea, cough, and indoor environment.

The methodological limitations of the present study are as follows. The study was not powered to find small differences in categorical outcomes. Hence, for the more rare symptoms, such as dyspnea grades 3 and 4, a large difference in prevalence was found but with large confidence intervals. There will be a substantial cost associated with

---

**TABLE 1. Telephone resurvey of previous postal respondents (n = 115), Oslo and Hordaland, Norway, 1998–1999**,†

<table>
<thead>
<tr>
<th>Prevalence (%)</th>
<th>Odds ratio</th>
<th>95% confidence interval</th>
<th>Kappa</th>
<th>Agreement (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Postal</strong></td>
<td><strong>Telephone</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morning cough</td>
<td>22.8</td>
<td>12.3</td>
<td>0.3</td>
<td>0.1, 0.8</td>
</tr>
<tr>
<td>Day cough</td>
<td>15.0</td>
<td>14.2</td>
<td>0.9</td>
<td>0.3, 2.8</td>
</tr>
<tr>
<td>Phlegm when coughing</td>
<td>18.9</td>
<td>17.1</td>
<td>0.8</td>
<td>0.2, 2.5</td>
</tr>
<tr>
<td>Cough of more than 3-month duration</td>
<td>14.5</td>
<td>10.9</td>
<td>0.3</td>
<td>0.0, 1.9</td>
</tr>
<tr>
<td>Dyspnea grade 1</td>
<td>13.3</td>
<td>18.6</td>
<td>3.0</td>
<td>0.7, 17.2</td>
</tr>
<tr>
<td>Dyspnea grade 2</td>
<td>12.3</td>
<td>14.9</td>
<td>1.4</td>
<td>0.5, 4.4</td>
</tr>
<tr>
<td>Dyspnea grade 3</td>
<td>3.5</td>
<td>1.8</td>
<td>0.3</td>
<td>0.0, 4.2</td>
</tr>
<tr>
<td>Dyspnea grade 4</td>
<td>3.5</td>
<td>0.9</td>
<td>0.0</td>
<td>0.0, 2.4</td>
</tr>
<tr>
<td>Attacks of breathlessness</td>
<td>15.8</td>
<td>17.5</td>
<td>1.3</td>
<td>0.4, 4.1</td>
</tr>
<tr>
<td>Wheezing</td>
<td>19.1</td>
<td>21.7</td>
<td>1.4</td>
<td>0.5, 4.4</td>
</tr>
<tr>
<td>Hay fever</td>
<td>20.7</td>
<td>22.0</td>
<td>1.3</td>
<td>0.2, 9.1</td>
</tr>
<tr>
<td>Eczema</td>
<td>17.3</td>
<td>18.7</td>
<td>1.3</td>
<td>0.2, 9.1</td>
</tr>
<tr>
<td>Mother has/had asthma</td>
<td>7.2</td>
<td>7.2</td>
<td>1.0</td>
<td>0.0, 78.5</td>
</tr>
<tr>
<td>Father has/had asthma</td>
<td>8.3</td>
<td>8.3</td>
<td>1.0</td>
<td>0.0, 78.5</td>
</tr>
<tr>
<td>Asthma diagnosis</td>
<td>5.7</td>
<td>6.6</td>
<td>N/A‡</td>
<td>N/A‡</td>
</tr>
<tr>
<td>Bronchitis diagnosis</td>
<td>21.1</td>
<td>24.2</td>
<td>1.8</td>
<td>0.4, 8.2</td>
</tr>
<tr>
<td>Emphysema diagnosis</td>
<td>2.1</td>
<td>2.1</td>
<td>N/A‡</td>
<td>N/A‡</td>
</tr>
<tr>
<td>Mother smoked when pregnant</td>
<td>8.6</td>
<td>9.7</td>
<td>0.94</td>
<td>0.6, 236.4</td>
</tr>
<tr>
<td>Younger than 7 years</td>
<td>25.0</td>
<td>28.8</td>
<td>5.0</td>
<td>0.6, 236.4</td>
</tr>
<tr>
<td>Moisture damage in home</td>
<td>10.7</td>
<td>12.5</td>
<td>1.4</td>
<td>0.4, 5.6</td>
</tr>
<tr>
<td>Floor deformaties in home</td>
<td>1.8</td>
<td>2.7</td>
<td>2.0</td>
<td>0.1, 118.0</td>
</tr>
<tr>
<td>Fungus smell in home</td>
<td>2.7</td>
<td>1.8</td>
<td>0.5</td>
<td>0.0, 9.6</td>
</tr>
<tr>
<td>Recent paint job in home</td>
<td>44.2</td>
<td>38.1</td>
<td>0.5</td>
<td>0.2, 1.3</td>
</tr>
<tr>
<td>Had dog at less than 2 years of age</td>
<td>15.3</td>
<td>17.3</td>
<td>3.0</td>
<td>0.2, 157.5</td>
</tr>
<tr>
<td>Had cat at less than 2 years of age</td>
<td>22.4</td>
<td>21.4</td>
<td>0.8</td>
<td>0.2, 3.7</td>
</tr>
<tr>
<td>Occupational exposure</td>
<td>38.0</td>
<td>44.4</td>
<td>2.2</td>
<td>0.8, 7.0</td>
</tr>
<tr>
<td>Symptoms while at work</td>
<td>13.5</td>
<td>13.5</td>
<td>1.0</td>
<td>0.3, 3.3</td>
</tr>
</tbody>
</table>

* Prevalence of binary questions in the two survey modes, with odds ratios between the survey modes, Cohen’s kappa, and agreement.
† The postal survey mode is the reference for odds ratios.
‡ N/A, not applicable because of one or more cells equal to zero in a 2 × 2 table.
discovering differences in the effect of survey method on outcomes with low prevalence. Similar considerations apply to the precision in kappa estimates.

Because of the time span between surveys, bias due to recall of previous answers should be unlikely. However, a real incidence or remission of symptoms may have influenced the results. Moreover, for questions that ask for events that may be remote, for example, ever wheeze, the event may be remembered at the first survey but forgotten at the second. Few data are available on 6–9 months' reliability and incidence and remission of respiratory symptoms.

Nonresponse bias is another problem. The responders to both survey modes represented only 45 percent of the target population (0.68 × 0.91 × 0.74). The response at different steps in the survey was variably dependent on age, gender, and region (data not shown). The response rates achieved here compare favorably with those of other surveys (21). A study of this type will always be susceptible to nonresponse bias, because of its sequential design. Even a 90 percent response at each stage (postal, telephone directory search, telephone response) would give a response rate of 72 percent.

Few epidemiologic studies have considered completeness of information between survey modes (item nonresponse). In our study, there was a clear decrease in missing information with the telephone survey mode, as has been found in a previous study (11). This may lead to differential nonresponse bias in comparisons between studies.

Reliability measurements for binary variables have limitations due to the scarcity of information contained in the 2 × 2 table from which they are derived (22). Thus, we chose to consider prevalence, kappa, and agreement.

For comparisons between different studies where one has used the postal mode and the other has used the telephone survey mode, our results are encouraging. Any differences in prevalence between studies are likely to be due to real differences, except for passive smoking, morning cough, and possibly questions on dyspnea. There appeared to be no systematic tendency for under- or overreporting of symptoms or diagnoses in the telephone survey relative to the postal survey. We speculate that the particularly large difference in morning cough between survey modes is related to its being the first question in the telephone interview. It is possible that respondents want to please the interviewer by providing positive answers, and this effect may be more pronounced for the first question. A social desirability bias may be involved for questions on passive smoking. Passive smoking may be seen as socially unwanted behavior, leading people to report less smoking to a telephone interviewer than on a relatively anonymous questionnaire. However, the evidence concerning desirability effects in general is equivocal (1).

For a mixed-mode health survey where postal and telephone surveys have been used, our results show a good comparability of survey methods. Agreement is good for most variables. For questions with a low prevalence, we encountered the problem that kappa is low while agreement is high (23), while for other rare exposures both kappa and agreement were high, even though prevalences were moderate to low for all these questions. Questions useful for studies of chronic obstructive pulmonary disease had low or intermediate reproducibilities (cough, phlegm, grades of dyspnea). Variables important to asthma and atopic disease epidemiology, such as asthma diagnosis and hay fever, had excellent reproducibilities, while wheezing and attacks of breathlessness were intermediate. Encouragingly, the questions on early life exposures and asthma in the family had high reproducibility, while indoor and work environment questions performed less well. Given the time spans involved, one would expect the opposite pattern. However, there were more missing answers to early life exposures than to environment questions, in line with expectations. In other words, early life exposures have good repeatability among those who feel they remember enough or are confident enough to answer, whereas environmental questions have a moderate repeatability all over.

For simple continuous and ordinal data such as height, weight, number of siblings, and education, comparability is excellent. Good repeatability of such factual questions is expected.

A recent study by Galobardes et al. (12) examines the impact of survey method on a respiratory health questionnaire, using a slightly different design. Their comparison of telephone versus postal survey modes involves subjects that were initial postal nonresponders but were then successfully surveyed by telephone and then attended a clinical examination where they completed a questionnaire. The only comparable question is wheezing. They report a kappa of 0.76 that is higher than ours. Another study of ad hoc clinical populations, using only a 2-week interval and the same method of administration, found a higher repeatability for wheeze questions (0.73–0.95 depending on study center) and a comparable repeatability for asthma diagnosis (24).

In conclusion, we found evidence of mode effects on completeness and prevalence. Symptoms of chronic bronchitis had moderate to low reproducibility, while questions relevant to asthma and atopic disease epidemiology had good reproducibility. Less morning cough and passive smoking at home and work were reported with the telephone survey mode. The survey mode may have a moderate effect on study results, depending on the specific questions asked.

ACKNOWLEDGMENTS

The study was supported by the Norwegian Research Council.

REFERENCES
