Problems and Proposals for Interview Data in Epidemiological Research

RALPH I HORWITZ* AND EUNICE C YU†

Horwitz R I (Yale University School of Medicine, Room IE-61 SHM, 333 Cedar Street, PO Box 3333, New Haven, Connecticut 06510, USA) and Yu E C. Problems and proposals for interview data in epidemiological research. International Journal of Epidemiology 1985, 14: 463–467.

We assessed the reliability of epidemiological data obtained by interview of 120 patients in a case-control study. The collected data, which were obtained by interview on two separate occasions, included such clinical and pharmaceutical features as history of lactation, hysterectomy, diabetes, type of menopause, and whether a woman had ever used exogenous oestrogens. Although we found generally high rates of agreement between interviews, errors in collecting and classifying data did occur, and were especially common for complicated clinical events, such as whether an oophorectomy accompanied the surgical removal of the uterus. Patients were also likely to disagree with previous responses when asked to recall a drug exposure occurring many years before. We identified seven sources of this variability, five in collecting the data, and two in coding. As a result of these findings, strategies are proposed for improving the quality of interview data obtained in epidemiological research.

Interview data obtained through questionnaires have gained widespread acceptance in epidemiology. This acceptance has occurred despite the recognition that individuals respond to questions in an interview by selecting a small amount of their total information to answer a specific question. Previous studies have demonstrated that some types of medical information are remembered more accurately than others. For example, patients remember details of illnesses requiring hospitalization better than illnesses treated in a physician’s office; and details of physician’s visits are better recalled than acute and chronic conditions managed by the patient alone. Variability in patient recall of past events can be anticipated, and at least two important determinants of the recall of medical information have been suggested. The first determinant is the interval between the clinical event and the interview: the longer the interval the less chance it will be remembered. A second determinant is the impact of the experience on the everyday life of the patient. The greater the impact the more likely the event will be recalled.

Data in epidemiological studies are usually obtained from entries in medical records, from direct conversation with patients, or from questionnaires answered by the patient. Recently, some studies have assessed the accuracy of interview data for certain items by comparing patient responses to medical record sources of information. The reliability of the data is seldom confirmed, however, by repeating the interview or medical chart extraction at a later date. Furthermore, the data are often obtained by excerpters or interviewers of varying training and ability, who are frequently aware of the patient’s status as diseased or non-diseased, and who frequently know the hypothesized causal relation under investigation. Since the epidemiological literature contains little discussion of these problems, we designed a study to assess the quality of data used in epidemiological case-control studies. We have already described our findings for data collected from hospital medical records. The investigation we are now reporting was intended: (1) to describe the occurrence and to identify sources of variability for data collected from direct patient interviews; and (2) to develop strategies for minimizing variation in data collection and coding.

METHODS

Data Technicians and Patient Interviews

Three data technicians, with similar training and experience, acted as interviewers in this study and interviewed all of the cases and controls. The data technicians had interviewed the patients initially as part of a large case-control study of the pharmaceutical antecedents of breast cancer among post-menopausal women. Interviews, which were carried out over the telephone, lasted 15–20 minutes and followed a
structured questionnaire format. After completing the interview, the information was transferred to a separate coding form that enabled the data to be used for computer analysis.

Interviews were completed on 542 post-menopausal women, aged 45 or older, evaluated at Yale-New Haven Hospital between 1 July 1976, and 30 June 1979. Cases in this sample included 129 women with newly diagnosed and histologically proven breast cancer. Three separate groups of controls, comprising 413 women without breast cancer were included: two groups of women chosen to be similar to each case for the diagnostic procedure (mammography or biopsy) used to evaluate the breast, age within four years, and date of the procedure; and a third group of women chosen from the medical and surgical wards of the hospital, and similar to the case in age, race, and date of hospitalization.

Selection of the Sample and Re-Interview by Data Technicians

We randomly selected 120 of the 542 patients for re-interview using the random numbers generated by the SAS (Statistical Analysis System) function UNIFORM. Among the 120 selected patients, 40 are cases and 80 are controls (40 from the diagnostic control groups and 40 from the hospital admission controls). The original interviews of these 120 patients were conducted between April 1980 and September 1981; the repeat interviews were conducted between December 1981 and February 1982. During the re-interview process, each data technician was blind to the findings and data from the first interview; and each subject was re-interviewed by a different data technician.

Data Collected for Analysis of Interview Variability

For the purposes of our analysis, we chose certain selected clinical and pharmaceutical features that were of potential importance to the analysis of the risk of breast cancer. Included clinical features were history of lactation, benign breast disease, bilateral oophorectomy, family history (first-degree relative) of breast cancer, previous hysterectomy, and hot flushes. The pharmaceutical feature we studied was whether the women had ever used oral oestrogens as post-menopausal replacement therapy.

Classifying Source of Variation

We were interested not only in measuring the amount of variation in collecting data from an interview, but also in identifying the reasons for the variation. Thus, after completing the interviews, we conducted detailed analyses of disagreements in the data and classified them into seven categories. Two of the categories could be considered as patient inconsistencies; three as interviewer errors; and two as occurring because of errors in data coding.

Disagreements occurring because of patient inconsistencies included:

(i) Patient disagreement: This occurred when the patient reported information in one interview that unambiguously contradicted data reported in the other interview. For example, a patient indicated on the first interview that she had taken oral oestrogens to relieve the symptoms of hot flushes. On the repeat interview, the patient again noted a history of hot flushes, but specifically denied ever using oral oestrogens.

(ii) Incomplete information: Occasionally, the patient reported partial information about a clinical event on one interview, but gave a complete explanation on the other interview. For example, in one interview the patient reported a syndrome of vasomotor instability with hot flushes, anxiety, and tachycardia. On the repeat interview, the patient reported that she had symptoms of anxiety occurring at the time her menstrual periods became irregular. Using the data obtained in the first interview, the patient would be classified as having a peri- or post-menopausal symptom complex that was a suitable indication for treatment with replacement oestrogens. Using the data from the repeat interview, the occurrence of the menopausal syndrome would be classified as uncertain.

Disagreements occurring because of interviewer error:

(i) Interviewer omission: This occurred rarely when an interviewer failed to ask a question that was included on the data questionnaire form.

(ii) Interview misinterpretation: Occurred when the interviewer incorrectly interpreted information reported by the patient. For example, when a woman reported receiving ‘a pink pill to treat my headaches and breast swelling’, the interviewer assumed that the medication could not have been replacement oestrogens, and did not pursue the response any further. Using physician record sources, we established that the medication was an oral oestrogen pill, and that the interview data were incomplete.

(iii) Patient disagreement and/or interviewer misinterpretation: Occasionally, we were unable to determine whether disagreements resulted from conflicting data reported by the patient or misinterpretations made by the interviewer. Rather than attribute the variation to one or the other source, we classified these errors under a separate distinctive category.
Disagreements occurring because of coding error:

(i) Coding error: Sometimes the patient reported consistent information, and the interviewer properly recorded the data on the interview questionnaire, but incorrectly coded the data. In these circumstances, there was a clear transcription error responsible for the disagreement.

(ii) Coding misinterpretation: Occasionally, both interviews result in similar data that did not clearly establish whether a particular attribute was present. In this circumstance, the data coder may have interpreted the data differently on two separate occasions. For example, a patient reported having post-menopausal hot flushes and receiving medication to treat the symptoms. However, the patient was unable to recall whether the medication was an oestrogen pill, some other ‘hormone’, or an anti-anxiety medication. Although the appropriate coding decision for oestrogen use in this circumstance is ‘uncertain’, the data technicians sometimes coded this information as ‘positive’ or as ‘negative’.

Statistical Procedures
The extent of agreement between interviews was assessed using per cent observed agreement for specific responses (positive and negative), and for per cent overall agreement. We also calculated an unweighted kappa statistic to provide an estimate of chance corrected agreement. The kappa statistic is calculated by the formula \( \kappa = (p_o - p_c)/(1 - p_c) \), where \( p_o \) is the observed proportion of agreement, and \( p_c \) is the proportion of agreement expected by chance. The value of kappa can vary from +1, indicating perfect agreement, to 0 indicating agreement no better than chance, and to -1 when agreement is less than expected by chance. Consistent with the recommendations of several research statisticians, the kappa value is interpreted as follows: the rating of agreement is poor when the value of the index is <0.4; the rating is fair when the index is 0.4–0.59; the rating is good when the index is 0.60–0.75; and the rating of agreement is excellent when the index value is >0.75.

RESULTS
Overall interview variability for eight selected clinical and pharmaceutical features is presented in Table 1. For each feature we present the total number of disagreements, per cent overall and specific agreement, and the value for kappa. The features are listed according to levels of overall agreement, and range from 97% for a history of hysterectomy or smoking to 83% for a history of hot flushes. The kappa statistic suggests good or excellent agreement for all eight variables.

In Table 2, we have classified sources of disagreement between interviews for five of the eight variables. There were a total of 58 disagreements for these features in the 120 studied interviews. Disagreements were more frequent for the variables hot flushes and oral oestrogens than for bilateral oophorectomy, family history of breast cancer, and smoking. Nearly two-thirds (65%) of the disagreements occurred because of conflicting patient reports, and an additional 17% resulted from errors in data coding. Infrequently, errors occurred due to interviewer misinterpretation or omission.

Table 3 reports disagreements between interviews for estimating age at occurrence of four selected clinical features. The number of disagreements ranged from 84 for estimating age at menopause, to 29 for age at oophorectomy. Perhaps most surprising was the large number of disagreements of two or more years. Although these were especially common for estimating age at menopause or oophorectomy, two year dis-

<table>
<thead>
<tr>
<th>TABLE 1</th>
<th>Overall interview variability for clinical and pharmaceutical features.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Features</td>
<td>Rate of agreement (N = 120 interviews)</td>
</tr>
<tr>
<td></td>
<td>Total no. disagreements</td>
</tr>
<tr>
<td>Hysterectomy</td>
<td>4</td>
</tr>
<tr>
<td>Smoking</td>
<td>4</td>
</tr>
<tr>
<td>Lactation</td>
<td>5</td>
</tr>
<tr>
<td>Family history of breast cancer*</td>
<td>6</td>
</tr>
<tr>
<td>Bilateral oophorectomy</td>
<td>11</td>
</tr>
<tr>
<td>Benign breast disease</td>
<td>17</td>
</tr>
<tr>
<td>Oral oestrogens</td>
<td>17</td>
</tr>
<tr>
<td>Hot flushes</td>
<td>20</td>
</tr>
</tbody>
</table>

* On the first or second interview, six patients were unable to recall whether there was a family history of breast cancer.
agreements also occurred when estimating age at menarche, for which the average age of onset was only 13 years.

Finally, we were especially interested in determining rates of agreement for such distinctly clinical characteristics as the reason a particular drug was started or stopped. In collecting data on drug use, we asked patients not only when a medication was begun, but also the clinical indication for drug use. Similarly, when asking patients to identify the date a medication was discontinued, we also inquired about the reason for discontinuing the treatment. In Table 4, we list the results of this analysis for use of oral oestrogens, and compare the agreement rates to those for the dose of oestrogen. Remarkably, we found similar agreement rates for all three features of drug use.

**DISCUSSION**

The results of this study suggest that patients can reliably recall and report much clinical data during interviews for epidemiological studies. For several variables (hysterectomy, lactation, smoking, and family history of breast cancer), agreement between interviews was excellent, with kappa values of 0.89–0.93. Overall, the rates of agreement for data collected by interview exceeded the agreement rates we have reported previously for repeated medical record extractions.

In the earlier study, overall agreement between medical record extractions was 81% for lactation, 84% for oral oestrogens, and 85% for hysterectomy, indicating agreement rates similar to those now being reported between interviews. However, for several other variables, interview rates of agreement were substantially higher than for the record extractions. For instance, agreement rates in record extractions were
80% for bilateral oophorectomy, and only 52% for benign breast disease (comparable interview rates were 91% and 86% respectively).

Since age at occurrence of certain clinical events is often studied as a risk factor for various cancers, we were especially interested in the unexpectedly high variability reported for these features. Perhaps most surprising was the large proportion of the disagreements that were for intervals at least two or more years apart. These data suggest that current evidence associating age at onset for these suspected risk factors for breast cancer may require re-evaluation.

Some investigators have argued that such clinical data as the reasons for starting or stopping pharmaceutical therapy (e.g., estrogen therapy) are too unreliable for use in epidemiological research. Our data dispute this contention. Patients were as consistent in reporting the reasons for starting or discontinuing drug use as they were for reporting the dose of drug used. This finding has special implications for studies that focus on the distinctive effects of clinical features on the association between risk factors (such as oral oestrogens) and the development of chronic disease (such as breast cancer).³

Methodological Implications

These observations have important implications for developing strategies to improve the basic quality of the data used in epidemiological research. Our data suggest that uncomplicated events and procedures, such as hysterectomy and family history of breast cancer, are remembered reliably. More complex procedures or clinical conditions appear to have more variability in patient recall. For example, although patients were consistent in their recall of the surgical removal of a uterus, they were often inconsistent in remembering whether the ovaries were also removed. Patients also were likely to disagree with themselves when asked to recall a drug exposure (e.g., oral oestrogens), that may have occurred 10 to 15 years before.

The substantial variability in recalling oestrogen exposure is particularly disturbing, since interview data about drug use occurring years before is collected frequently in epidemiological research. The variability described in this study confirms earlier research in which the recall variability increased as the interval between exposure and the interview was increased. Our findings suggest that interview data may be best suited for uncomplicated clinical events or procedures and for current or recent drug exposure. For more complicated medical procedures and for drug exposures occurring years before, variability in patient recall may make interview data unreliable. In these circumstances, investigators may need to emphasize the use of physician or hospital medical records.

The data collected in our study was part of a case-control study of breast cancer, and we cannot be certain that the results are applicable to epidemiological studies of other disorders. However, the clinical and methodological considerations for breast cancer are typical of those for other cancers and chronic diseases, and the exposure variables included in our study are standard epidemiological risk factors.

Our data also emphasize the need to simplify coding criteria and to monitor their application. This latter requirement cannot be readily achieved if data are collected during interviews using the increasingly popular precoded data forms in which information is collected and coded in a single step. We believe these forms should be discouraged, since by combining two steps in obtaining interview data, the precoded forms will make it more difficult to monitor and assess adequately the reliability of each individual step. We recognize that this proposal will be unpopular with epidemiologists who know that precoded forms are efficient in saving data technicians' time and effort. However, because we have demonstrated that errors can occur both during data collection and coding, epidemiologists need strategies that enhance the quality of the information at each step. Until better methods, including those suggested in this paper, have been developed and tested, data collection and coding strategies should be simple and easily monitored.

REFERENCES

2 Madow W G. Interview data on chronic conditions compared with information derived from medical records. Vital Health Statistics, Series 2, No 23, 1967.