Clinical studies using medical record review should include careful training and quality assurance methods to enhance the reliability and validity of data obtained from the records. Because of time and budget constraints, comprehensive assessments of data quality and reliability, including masking of medical record abstractors, are not always possible. This paper describes the abstractor training and quality control methods and results of a masked medical record review study. The medical record review study was carried out within a larger multisite study of the effectiveness of screening mammography in preventing breast cancer mortality with an observation period within 1983 and 1993, with mortality follow-up through 1998. An eight-step program was developed to train medical record abstractors and monitor the quality of their work. A key follow-up component to the training protocol was a 5% reabstraction of medical records (n = 160), masked and reviewed by a second abstractor. High agreement was found between initial (unmasked) abstractors and masked abstractors for all key exposure variables (kappa ranged from 0.76 to 0.91), with no evidence of biased directionality by unmasked reviewers. Rigorous ongoing training programs for medical record abstractors provide assurance of good quality control in large multisite studies. Additionally, a masking study with a subsample of subjects may be a feasible and cost-effective alternative to the time- and cost-intensive methodological approach of masking all medical records.

Abbreviation: HMO, health maintenance organization.

Strategies for increasing the reliability, validity, and quality of data collected from medical record reviews include thoroughly training medical record abstractors, masking (i.e., blinding) abstractors to study hypotheses and case/control assignment, and assessing interrater reliability and agreement (1–3). In many studies, however, the cost and logistical challenges required to successfully incorporate these strategies exceed the project’s timeline and available budget. We found surprisingly few data in the scientific literature on the topics of masking and medical record abstractor training. With respect to masking, one challenge is that neither “mask” nor “blind” is an ideal Medical Subject Headings term from an epidemiologic standpoint, because the term “mask” refers to a device covering the nose/mouth and “blind” refers to a visually impaired person. Therefore, it is difficult to ascertain how often masking of medical record reviews is performed or studied. We also had difficulty finding an appropriate Medical Subject Headings term for “abstractor training,” making it difficult to determine how investigators are training medical record abstractors or how
often an established or thorough training protocol is followed.

One published review of 244 articles using medical record abstractions found that only 18 percent mentioned abstractor training, only 3 percent mentioned abstractor masking to study hypotheses and patient assignment, only 5 percent mentioned interrater reliability, and less than 1 percent statistically tested interrater agreement (4). The findings of this review could be interpreted as a failure of researchers to adequately design retrospective medical record studies or simply as a failure of these researchers to report their methods in detail.

In this paper, we describe the methods of a rigorous abstractor training protocol and the findings of a masked medical record review sub-study carried out within a large multisite case-control study of the effectiveness of screening mammography in preventing breast cancer mortality.

MATERIALS AND METHODS

Study population

This study was performed under the auspices of the Cancer Research Network. The Cancer Research Network consists of the research programs, enrollee populations, and databases of 10 health maintenance organizations (HMOs) that are members of the HMO Research Network. Six of the 10 health care delivery systems in the network participated in this study: Group Health Cooperative, Harvard Pilgrim Health Care, HealthPartners Research Foundation, and three regional divisions of Kaiser Permanente: Northwest (Oregon), Northern California, and Southern California.

The overall goal of the Cancer Research Network is to increase the effectiveness of preventive, curative, and supportive interventions that span the natural history of major cancers among diverse populations and health systems, through a program of collaborative research. This overarching aim, coupled with the expertise of the investigative team and the geographically dispersed population base, fosters efficient and effective research in cancer prevention, early detection, and treatment.

Study design

A case-control study was conducted to examine the efficacy of breast cancer screening among women in two age cohorts at two different breast cancer risk levels. Data collection consisted of medical record reviews, using a computerized abstraction form, carried out at six HMOs located in five states (Massachusetts, Minnesota, Washington, Oregon, and California). Institutional review boards approved of the study procedures at the individual sites. Eligible cases were women who died of breast cancer and were aged 40–67 years at the time of diagnosis (n = 1,351). Eligible controls were women without breast cancer who were matched to cases by HMO site, age, and breast cancer risk level (n = 2,501). (Sample numbers in the main study are currently being analyzed and are subject to change.) Breast cancer risk level was categorized as increased if a family history of breast cancer and/or a history of breast biopsy was noted; otherwise the risk level was considered average. The study window for breast cancer diagnosis was January 1, 1983, through December 31, 1993; follow-up extended to December 31, 1998, for ascertainment of deaths due to breast cancer. Medical record data were reviewed for a 3-year period prior to the date of first knowledge or symptom of breast cancer (the index date) for cancer screening variables. For the masking substudy, a sample of approximately 5 percent (n = 160) of the main study population, including both eligible cases and controls, was used at five of the six participating sites. The initial medical record abstractor was not aware of which medical records would be selected later for the 5 percent review.

Abstractor training and quality control

For the main case-control study, a standardized protocol was used to train medical record abstractors and to abstract clinical data from medical records. The six site coordinators and 15 medical record abstractors received extensive training, including the following.

Training manual. As the first step in training, an educational training manual was distributed to all sites for review. The manual included a study overview and information on procedures for abstraction, quality assessment, progress reports with which to track data collection, an abstraction form and coding instructions, a quick reference sheet on all variables, a glossary of terms, a summary of literature relevant to key study variables, and the standardized training examples (see below).

Standardized training examples. Key study variables that would be important to the main study and potentially challenging for abstractors were determined a priori. These variables included the index date, information on screening and diagnostic mammograms, and the cause of death. When pilot-testing the abstraction form for the main study, several examples were pulled from medical records and organized into training examples. Six training examples were prepared for each of these variables, where abstractors would read an overall summary of the medical chart pages, peruse the actual photocopied chart pages, and complete an abstraction training form for each.

Individual orientation. Two or more individual orientation sessions were arranged with each site’s abstraction team and the Lead Investigator (J. G. E.) and/or Lead Study Coordinator (L. M. R.) via conference call. At the first session, the training manual was reviewed and site-specific issues were addressed. At the second session, the training examples were reviewed. Additional sessions were scheduled if deemed necessary by the site coordinators.

Double-review of the first 10 charts abstracted. The first 10 charts abstracted were double-reviewed by another site abstractor. The two abstractors then met to go over discrepancies in abstraction. All discrepancies were logged onto an audit adjudication form and sent to the Lead Study Coordinator for tracking of areas of difficulty.

On-site visits. After the first 10 charts were double-reviewed, individual on-site visits were made by the Lead Investigator and/or the Lead Study Coordinator. These per-
sons reviewed the initial chart audits and answered abstraction questions posed by the site abstractors. **Monthly double-review.** Each month, a double-review of one eligible case or control and one ineligible case or control was performed for each abstractor. The charts were selected and prepared by site coordinators, and abstractors met to discuss discrepancies. An audit adjudication form was completed on all discrepancies and was sent to the Lead Study Coordinator.

**Twice-monthly conference calls.** Conference calls conducted twice monthly included an ongoing training component in which examples of difficult chart abstractions were distributed prior to the call for all abstractors to review, discuss, and reach consensus on. The Lead Study Coordinator also discussed areas of difficulty noted in the monthly double-reviews. A decision log (containing a summary of the discussions held and decisions made during the conference calls) was updated after each call and disseminated to the group to keep in their training manuals.

**Simultaneous data collection and cleaning.** Abstractors entered medical record data directly into laptop computers using Microsoft Access (Microsoft Corporation, Seattle, Washington) and sent data each month to the lead coordinating site, where abstractions were tracked and data were cleaned. Each month, site coordinators were asked to look into possible errors flagged during data cleaning. Site coordinators and abstractors either corrected errors or sent back information on why the original information was correct. This work was all done before the next monthly file transfer of data; the data cleaning process was repeated monthly. Therefore, data cleaning occurred continuously throughout the chart review process. This simultaneous data cleaning method served as another ongoing training opportunity for abstractors, so that errors were not repeated throughout the course of chart abstractions.

**Logistics of the masking substudy**

The subject’s case/control status was not hidden from abstractors in the main study, whereas abstractors for the masking study did not know case/control status. Abstractors in both the main study and the masking substudy were the same, but for each medical record the abstractor for the masking substudy was different from the initial abstractor in the main study. The masking study focused on the 3-year window prior to the first knowledge or symptom of breast cancer (i.e., the index date) to review for history of mammograms, including descriptions of the examinations and examination results. Information on the index date was not included in the masking substudy, since it could have alerted the abstractors to the subject’s case/control status. A shortened version of the main study abstraction form was used in the masking substudy.

Two methods were used to prepare the medical records for masked review: full-record and partial-record. At three sites, project coordinators prepared the medical records by masking pages that could reveal case/control status and provided the “full record” to the masked reviewer. At each site, appropriate methods of masking the medical records were devised, using some combination of paper clips, adhesive notes, and white sheets of paper (e.g., in situations where the front of the chart was stamped “deceased,” this was covered with paper). Two sites chose to use a “partial record” masking procedure by photocopying relevant pages in the 3-year window for abstraction by the masked abstractor.

**Definition of mammography variables**

Abstractor agreement was assessed for four key mammography variables: 1) total number of mammograms, 2) total number of screening mammograms, 3) total number of diagnostic mammograms, and 4) classification of all mammograms. Total number of mammograms was the count of any mammogram noted in the 3-year data collection window, regardless of the reason or results. Total number of screening mammograms included all mammograms in the study window for which the reason was noted to be “patient possibly symptomatic,” “clear-cut symptoms,” or “other positive test.” Classification of all mammograms was defined for each woman in terms of whether she had had all screening mammograms, all diagnostic mammograms, both screening and diagnostic mammograms, or no mammograms.

**Analytical plan**

The kappa (κ) statistic (5) was used to determine the level of agreement for each mammography variable between masked and nonmasked chart reviews. The weighted kappa statistic was used for all comparisons except classification of all mammograms, for which a simple or nonweighted kappa statistic was used. The following guidelines were used to interpret the strength of agreement when kappa was positive: \( \kappa \geq 0.61 \) was considered to represent substantial agreement; \( \kappa \geq 0.81 \) was considered to represent almost perfect agreement (6). Bowker’s test of symmetry (7) was used to evaluate possible biased directionality in the case of disagreement between masked and nonmasked reviews. Separate analyses were performed to search for possible differences in abstractor agreement by case/control status, as well as by the masking modality employed (i.e., full or partial records).

**RESULTS**

In the masking substudy, 160 charts were reviewed, with a range of 29 to 38 charts at each of the five participating sites. The charts were from 79 cases and 81 controls, with 97 women at average risk for breast cancer and 63 women at increased risk for breast cancer. The women were aged 40–65 years at the index date.

Agreement between the masked and unmasked abstractors was high for all mammography variables, with kappa statistics ranging from 0.76 to 0.91 (figure 1 and table 1). Bowker’s test of symmetry revealed no biased directionality or systematic pattern of disagreement between masked and nonmasked reviews. Similarly, kappa statistics were very high and were similar for cases and controls (table 1). Of the 160 charts, 101 were abstracted using the
full-record masking method (three sites) and 59 were abstracted using the partial-record masking method (two sites). For total number of mammograms, number of screening mammograms, and overall mammogram classification, agreement for each masking method was almost perfect (table 1). For number of diagnostic mammograms, agreement was slightly lower for cases than for controls and was lower with the full-record method than with the

**TABLE 1. Agreement between masked medical record abstractors and initial nonmasked abstractors in a study of screening mammography and breast cancer risk (n = 160), Cancer Research Network, 1983–1993**

<table>
<thead>
<tr>
<th></th>
<th>All mammograms</th>
<th>Total no. of mammograms per woman</th>
<th>Diagnostic mammograms</th>
<th>Classification of all mammograms†</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Weighted (\kappa)</td>
<td>95% CI*</td>
<td>Test of symmetry (p value)</td>
<td>Weighted (\kappa)</td>
</tr>
<tr>
<td>All women (n = 160)</td>
<td>0.91</td>
<td>0.86, 0.96</td>
<td>0.52</td>
<td>0.82</td>
</tr>
<tr>
<td>Stratified by case/</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>control status</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case (n = 79)</td>
<td>0.94</td>
<td>0.88, 0.99</td>
<td>0.68</td>
<td>0.82</td>
</tr>
<tr>
<td>Control (n = 81)</td>
<td>0.88</td>
<td>0.81, 0.96</td>
<td>0.78</td>
<td>0.81</td>
</tr>
<tr>
<td>Stratified by method</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full record (n = 101)</td>
<td>0.89</td>
<td>0.82, 0.95</td>
<td>0.68</td>
<td>0.81</td>
</tr>
<tr>
<td>Partial record (n = 59)</td>
<td>0.96</td>
<td>0.90, 1.00</td>
<td>0.92</td>
<td>0.83</td>
</tr>
</tbody>
</table>

* CI, confidence interval.
† Simple kappa was used instead of weighted kappa.
partial-record method; however, these differences were not statistically significant.

**DISCUSSION**

Observational studies have an important place in clinical research. We must often rely on data obtained from clinical records, because randomized clinical trials would be either too costly, too time-consuming, or unethical. For epidemiologic studies of rare clinical disorders, such as death from a specific cancer, multisite research projects are often required in order to amass an adequate sample size for testing the study hypotheses. In the present study, it is very likely that careful training played a role in the results of the substudy, which showed almost perfect agreement among abstractors.

An eight-step medical record abstractor training program was developed and carried out in this multisite study. The educational training manual containing information about the study and the study protocol provided the basic foundation for the abstractors. Standardized training examples provided a “test” situation in which abstractors could practice implementing their understanding of key areas of the abstraction form and coding instructions. The individual orientation sessions were a critical step toward ensuring abstractor comprehension of key variables, in addition to working through site-specific logistical concerns. Double-review of the first 10 charts abstracted, in addition to the required monthly double-review, allowed site abstractors to work through some charts together and reach consensus on abstraction issues.

In retrospect, it was invaluable to have site staff meet study leaders during on-site visits. These visits increased morale and rapport among study staff and provided insight about the vast differences in clinical care systems and medical charts among sites. The regular, twice-monthly conference calls gave us an opportunity to continually reach consensus on difficult chart abstractions from the sites. Finally, simultaneous data collection and cleaning was an invaluable step toward preventing repeated abstraction errors throughout the course of the study.

The benefits of this extensive training program are demonstrated by the masking study results. We found almost perfect agreement between unmasked and masked medical record abstraction on assessment of screening and diagnostic mammography and no evidence of biased directionality. For our primary variable of interest, screening mammograms, abstractor agreement was high and was consistent for cases and controls and for the two masking methods (either full or partial medical record review).

Our study was able to provide an overall estimate of agreement within 10 percent, with slightly less agreement noted for the stratified analyses. With almost perfect levels of agreement, it is difficult to assess the presence of bias. Bias in a 2 × 2 table can be assessed by comparing the off-diagonal elements to see whether they are distributed equally in both directions. When kappa is high, the total number of off-diagonal elements is very small. In such cases, typically, the overall sample size would have to be in the thousands for investigators to have sufficient power to detect bias when the overall level of agreement is high, because most observations are not included when evaluating bias. Therefore, although our masking study lacked the power to definitively rule out abstractor bias, any contribution would be small because of the very high agreement.

Performing 100 percent double chart review for all subjects in the main study, with the second abstractor “masked” or “blinded” as to case/control status, would have been extremely costly and labor-intensive. In the main study, we reviewed nearly 8,000 medical records to assess eligibility status at six clinical sites. Each site had different types of medical records and clinical systems. For example, at one site it was possible for an individual woman to be seen at several different outpatient centers over a few years, each with its own medical record. This site did not have centralized medical records; thus, the chart abstractor would be required to travel to multiple sites, with distances between centers of up to 150 miles (240 km), to complete chart abstractions. A masked review of all potential cases and controls for our main study would have doubled the time and budget needed for study abstractors and coordinators. By conducting a masked review of a sample, we were able to evaluate the reliability of the record abstraction process for the overall study.

In the masking substudy, we considered asking each masked abstractor to guess the case/control status of each woman. We decided against this approach, since we did not want the abstractors thinking about this review as a “test” and actively looking for clues about the subject’s case/control status. It is possible that a masked abstractor could have reviewed the charts in a different manner than the initial abstractor because he or she saw the work as a “test.” A similar case-control study conducted by Group Health Cooperative asked masked abstractors to guess the actual case/control status of each woman (breast cancer case vs. control); the masked abstractors correctly identified case/control status for 74 percent of the subjects (n = 86; William Barlow, Group Health Cooperative, personal communication, 2002).

Additionally, it was not possible to emphasize all variables in the training efforts and the masking substudy, simply because there were too many variables on the full abstraction form. Therefore, it is difficult to know whether or not focusing on key variables meant that abstractors did well on key study variables and less well on others. We feel, however, that while the training examples and the masking substudy emphasized a few key variables, the monthly ongoing training and reabstraction emphasized quality on all facets of the abstraction form, hopefully diminishing the possibility of differential attention.

In summary, this masking substudy provided assurance that abstractors were reliably assessing key study variables. The almost-perfect agreement noted in the present study between masked and unmasked medical record abstractors, based on a 5 percent sample of cases and controls, supports the benefits of standardized training of abstractors and rigorous quality assurance measures. In multisite medical record abstraction studies, it is critical to develop thorough standardized training and quality assurance approaches, with consideration of time and budget constraints.

ACKNOWLEDGMENTS

This project was supported by grant CA79689 from the National Cancer Institute. The overall Principal Investigator for the Cancer Research Network is Dr. Edward H. Wagner.

The authors thank the project coordinators at each site for their hard work: Sherry Falls, Sharon Flores, Ana Macedo, Jill Mesa, Deborah Reck, Vanessa Ryan, and Carmen West. The authors appreciate the dedication of Cary Williams and the entire project staff, especially the abstractors.

REFERENCES